IMDB reviews: Analyzing and predicting emotions

Presented by Artur Oganesyan 47496

1. Introduction

Business problem

Online movie reviews play a significant role in shaping audience perceptions and influencing future viewership. IMDb, being one of the largest platforms for movie reviews, provides a rich source of user-generated content that can be analyzed to gain insights into audience sentiment. Idea is to analyze 50000 reviews and classify these reviews into different categories. This project is aimed to extract valuable information from IMDb reviews to understand the public's opinion on movies, identify common themes, and build a sentiment classification model.

Key Project Questions:

- What are the primary sentiments expressed in IMDb reviews?
- Is there a correlation between sentiment scores and IMDb ratings?
- What are the most commonly used words in the reviews?

Objectives

- Perform sentiment analysis on IMDb reviews.
- Identify key themes in user reviews using Natural Language Processing (NLP) techniques.
- Create a custom emotion dictionary.
- Build and evaluate a sentiment classification model based on created custom dictionary.
- Compare my model with an existing one (VADER).
- Predict new reviews based on new model.
- Provide insights and recommendations based on the analysis.

2. Dataset Overview

Description of Dataset

IMDB dataset has 50000 movie reviews for natural language processing or Text analytics.

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. So, this dataset is quite suitable in order to predict the number of positive and negative reviews using Natural Language Processing (NLP) as it contains user-generated text, which can be preprocessed and analyzed for sentiment.
The dataset contains 50000 movie reviews and consists of two columns: review and sentiment.

Source of Data & Method of Collection

After analyzing many different sources, I have decided to choose the dataset that had been taken from Kaggle: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

3. Analytical Methods

Data Cleaning

- Removal of missing values.
- Elimination of URLs, HTML tags, special characters, and non-alphanumeric symbols.
- Conversion of text to lowercase.
- Removal of stopwords using the NLTK library.
- Tokenization of text data.

Descriptive Analysis

To understand the nature of the dataset, several descriptive methods were applied:

- Frequency of the top 20 most common words in the dataset.
- Bar chart visualization of the most common words using Seaborn.
- Word cloud generation to visualize popular terms in the dataset.

Sentiment Analysis with VADER

The **VADER** tool was employed to predict sentiment polarity. Each review was classified as **positive** or **negative** based on the compound sentiment score provided by VADER.

Custom Emotion Classification

A custom dictionary was developed to classify reviews into four emotion categories:

- Excitement
- Anger
- Sadness
- Neutral

The classification was performed by counting the occurrences of predefined keywords in each review.

Machine Learning Model for Emotion Prediction

A Logistic Regression model was built to predict emotion categories. The workflow included:

- Text vectorization using TF-IDF (Term Frequency-Inverse Document Frequency).
- Splitting the dataset into training and testing sets.
- Model training.
- Performance evaluation using accuracy, precision, recall, and f1-score.

Comparative Analysis

A comparison was made between the ML model and the VADER sentiment analysis. It was conducted in order to observe the difference between VADER model and the model that I had created.

Subjectivity Analysis

With the help of TextBlob, I have conducted a subjectivity analysis which assigns a subjectivity score ranging from 0 (objective) to 1 (subjective).

Prediction of New Reviews

I also added a custom function that was created to predict both emotion category and subjectivity for new, unseen reviews.

4. Analysis of Data

In this section, I will present a detailed analysis of the obtained data that I have got from my work. I will show your results and try to explain it.

Showcasing Examples of Reviews

Here I will show you how I handled with data cleaning, and examples of positive and negative reviews in order to see how data cleaning works.

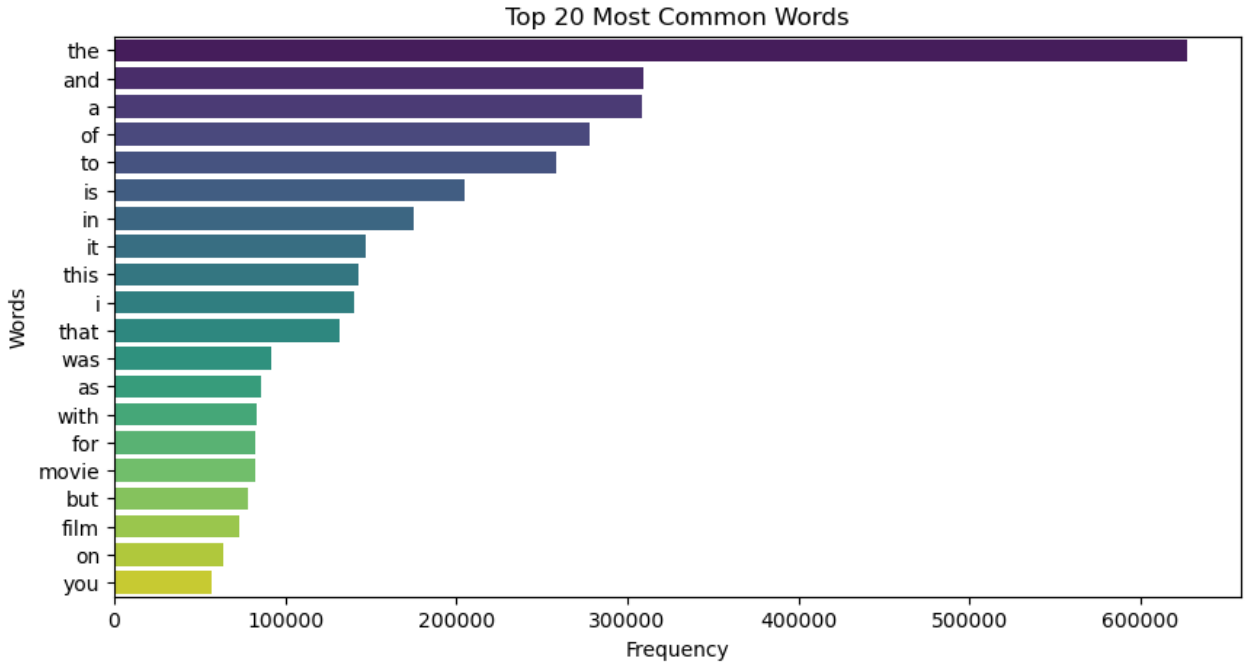
```
Positive Review (Cleaned):
one of the other reviewers has mentioned that after watching just 1 oz episode youll be hooked they are right as this is exactl
y what happened with methe first thing that struck me about oz was its brutality and unflinching scenes of violence which set i
n right from the word go trust me this is not a show for the faint hearted or timid this show pulls no punches with regards to
drugs sex or violence its is hardcore in the classic use of the wordit is called oz as that is the nickname given to the oswald
maximum security state penitentary it focuses mainly on emerald city an experimental section of the prison where all the cells
have glass fronts and face inwards so privacy is not high on the agenda em city is home to manyaryans muslims gangstas latinos
christians italians irish and moreso scuffles death stares dodgy dealings and shady agreements are never far awayi would say th
e main appeal of the show is due to the fact that it goes where other shows wouldnt dare forget pretty pictures painted for mai
nstream audiences forget charm forget romanceoz doesnt mess around the first episode i ever saw struck me as so nasty it was su
rreal i couldnt say i was ready for it but as i watched more i developed a taste for oz and got accustomed to the high levels o
f graphic violence not just violence but injustice  watching oz you may become comfortable with what is uncomfortable viewingth
ats if you can get in touch with your darker side

Negative Review (Cleaned):
basically theres a family where a little boy  thinks theres a zombie in his closet  his parents are fighting all the timethis m
ovie is slower than a soap opera and suddenly jake decides to become rambo and kill the zombieok first of all when youre going
to make a film you must decide if its a thriller or a drama as a drama the movie is watchable parents are divorcing  arguing li
ke in real life and then we have jake with his closet which totally ruins all the film i expected to see a boogeyman similar mo
vie and instead i watched a drama with some meaningless thriller spots3 out of 10 just for the well playing parents  descent di
alogs as for the shots with jake just ignore them
```

	review	sentiment	cleaned_review
0	One of the other reviewers has mentioned that ...	positive	one of the other reviewers has mentioned that ...
1	A wonderful little production. The...	positive	a wonderful little production the filming tech...
2	I thought this was a wonderful way to spend ti...	positive	i thought this was a wonderful way to spend ti...
3	Basically there's a family where a little boy ...	negative	basically theres a family where a little boy ...
4	Petter Mattei's "Love in the Time of Money" is...	positive	petter matteis love in the time of money is a ...

You can see that all special characters have been removed so the data is ready for analysis.

Bar Chart of Top 20 Most Common Words



The

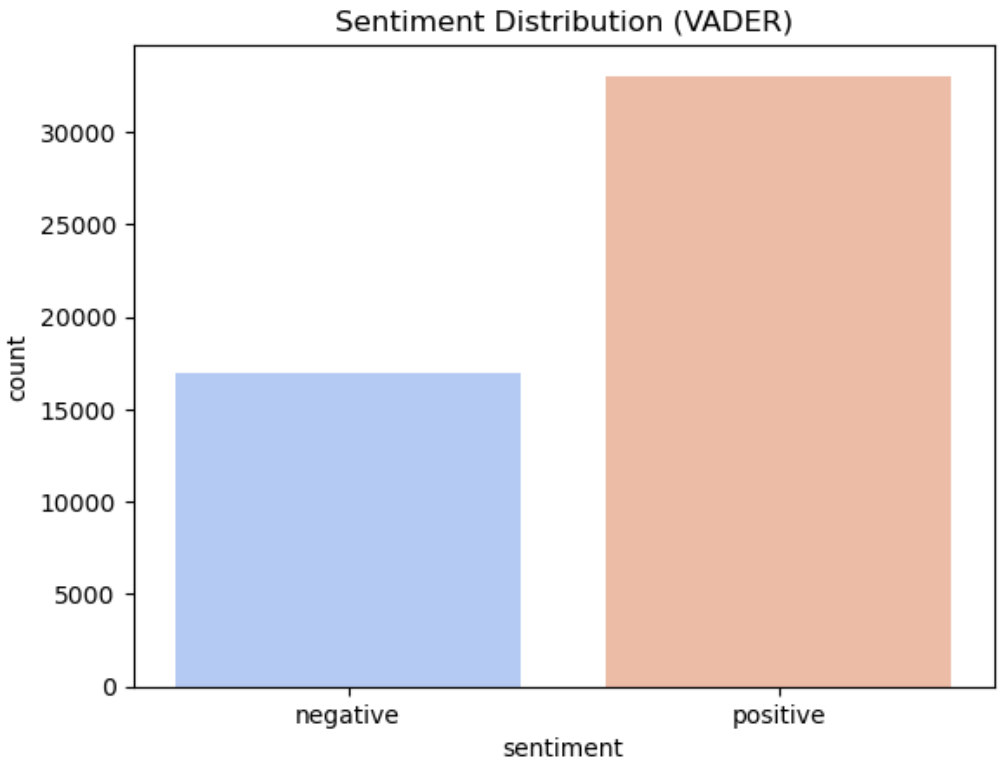
This bar chart shows the **20 most frequently used words** in the dataset. These words give insight into the overall language structure of the reviews. The most common words are **"the," "and," "a,"** which are **stopwords** (common words with little meaning). Besides stopwords, relevant words like **"movie," "film,"** and **"character"** appear frequently, indicating that the reviews revolve around movies, their

stories, and characters.

Word Cloud for Reviews

The word cloud visualizes the **most frequently used words** in a more aesthetic way, where larger words appear more frequently. Words like **"movie," "film," "character," "story," "see,"** and **"good"** are prominent, indicating that users often discuss the movie's storyline, characters, and quality. Positive terms like **"good," "great," "love,"** and **"watch"** show that many reviews are positive in nature.

Interpretation of Sentiment Distribution (VADER)



VADER chart shows the number of **positive** and **negative** reviews in the dataset after applying the VADER sentiment analysis. The chart indicates that there are more positive reviews than negative ones. You can see that there are more than 30000 positive and almost 17500 negative reviews. VADER classifies text based on sentiment intensity, where a positive score means the review expresses positive emotions, and a negative score indicates negative emotions.

Model Performance Evaluation

Accuracy: 0.9229

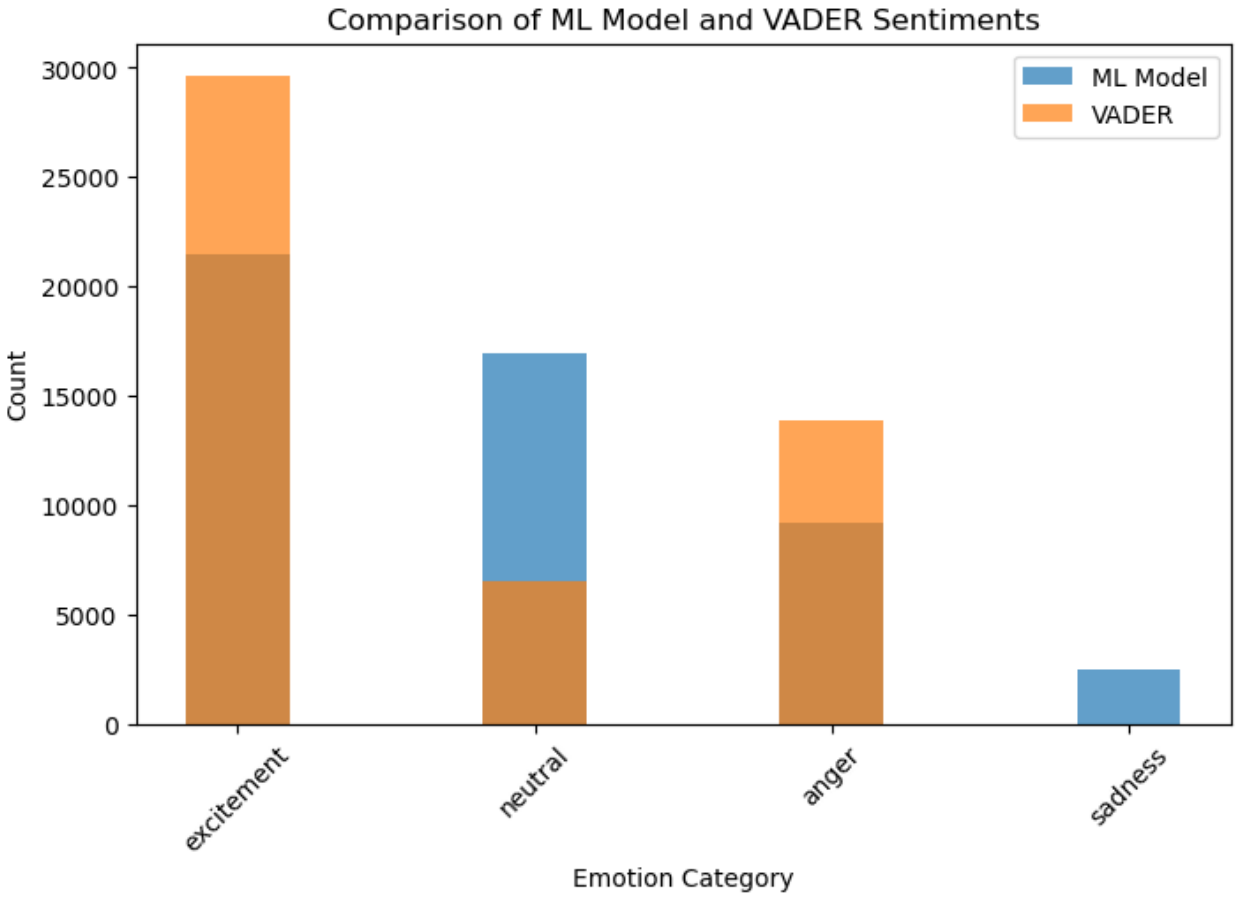
	precision	recall	f1-score	support
anger	0.96	0.85	0.90	1835
excitement	0.96	0.96	0.96	4299
neutral	0.86	0.99	0.92	3370
sadness	0.97	0.44	0.61	496
accuracy			0.92	10000
macro avg	0.94	0.81	0.85	10000
weighted avg	0.93	0.92	0.92	10000

My model demonstrates good performance with an accuracy of **92.29%**, effectively classifying sentiments across four categories. It performs exceptionally well in detecting excitement (96% F1-score) and neutral (92% F1-score) sentiments, indicating consistent reliability. The model also excels in precision for anger (0.96) but struggles with recall (0.85), missing some anger reviews. However, the model shows weak performance in detecting sadness, with a low recall of 0.44, suggesting that many sadness reviews go undetected. The overall precision and weighted F1-score of 0.92 highlight the model's confidence and balanced classification.

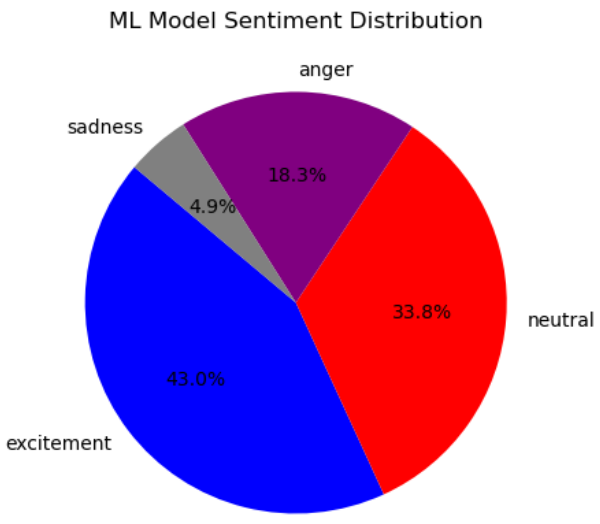
Model performs great on excitement and neutral emotions. Sadness needs improvement — the model finds sadness when it predicts it, but it's not sensitive enough to catch all the sad examples.

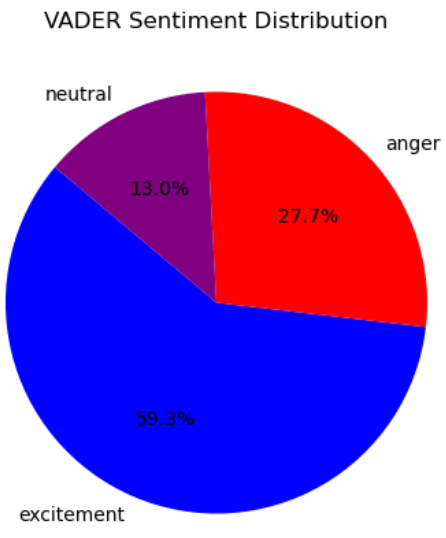
My ML Model VS VADER Comparison

In order to highlight the difference between VADER and the model that I had created, I decided to check the results of my model and VADER sentiments. The results you could see below:



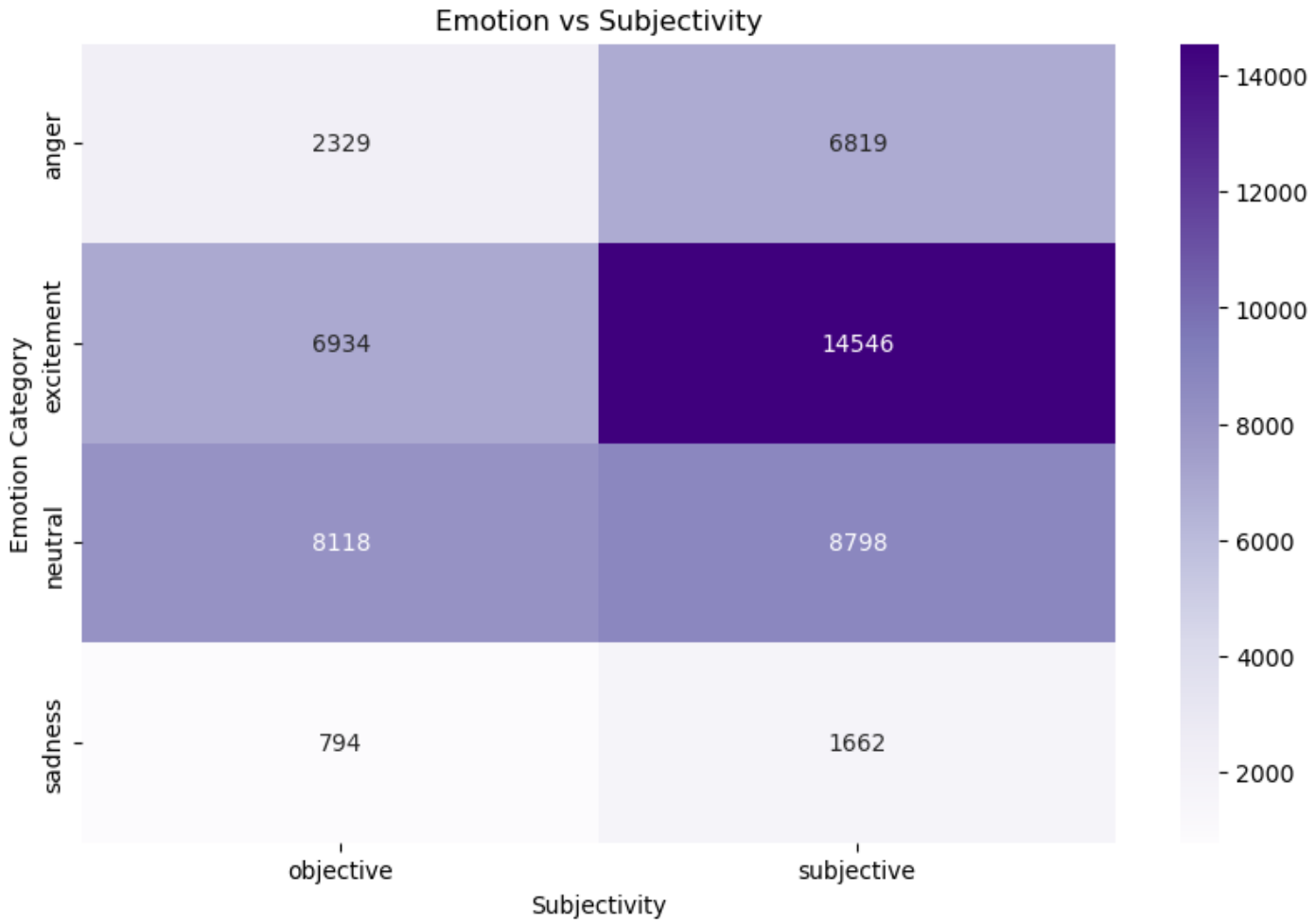
The first image presents a bar chart comparing sentiment classifications between a machine learning (ML) model and VADER. The results indicate that VADER assigns significantly more reviews to the "excitement" category compared to the ML model, while the ML model classifies a larger portion of reviews as "neutral." Both models show similar trends for "anger," though VADER categorizes slightly more reviews under this emotion. Interestingly, the ML model is the only one assigning a notable count to the "sadness" category, suggesting a difference in sentiment detection approaches.





The second image consists of two pie charts illustrating the sentiment distribution of each model. The left pie chart, representing the ML model, shows a balanced distribution, with "excitement" (43%) and "neutral" (33.8%) as the most frequent sentiments, followed by "anger" (18.3%) and "sadness" (4.9%). In contrast, the right pie chart, representing VADER, heavily favors "excitement" (59.3%) and "anger" (27.7%), while "neutral" (13%) is much less common. These differences suggest that VADER tends to classify emotions more intensely, whereas the ML model takes a more moderated approach, distributing reviews more evenly across categories.

Subjectivity Analysis



The image presents a heatmap comparing emotions (anger, excitement, neutral, sadness) with subjectivity (objective vs. subjective). The color intensity represents the frequency of occurrences, with darker shades indicating higher values. The data shows that "excitement" has the highest number of subjective reviews (14,546), followed by "neutral" (8,798) and "anger" (6,819). Objective reviews are more evenly distributed across categories, with "neutral" being the most common (8,118) and "sadness" the least frequent (794). This suggests that excitement and anger tend to be more subjective, while neutral sentiments are more evenly split between objective and subjective reviews.

Predict Emotion and Subjectivity for New Reviews

You can see that I have added new review for testing, in this case you can see that "the movie was incredibly thrilling" is considered as excitement sentiment with predicted subjectivity: “subjective”.

```
Predicted Emotion: excitement
Predicted Subjectivity: subjective

'The movie was incredibly thrilling, with an amazing storyline and stunning visuals!'
```

Subjectivity plays a crucial role in sentiment analysis because it helps distinguish between **opinions and facts** in textual data. Understanding subjectivity improves the accuracy and depth of sentiment classification, making models more context-aware. Now you can see that there is an opportunity to predict not only emotion for future reviews but also subjectivity.

5. Description, Strategies, and Conclusion

Description of a problem

The analysis explores sentiment classification in IMDB movie reviews, comparing results between a machine learning (ML) model and VADER. The study investigates the distribution of sentiments across different emotion categories (excitement, anger, neutral, and sadness) and further examines the relationship between sentiment and subjectivity. Various visualizations, including bar charts, pie charts, and heatmaps, help illustrate differences in classification trends between models. The findings highlight how each model interprets sentiment, with VADER tending to assign more extreme emotions, while the ML model provides a more balanced distribution.

Recommended Strategies

- **Preprocessing Techniques** – Cleaning the dataset by removing stopwords, punctuation, and special characters to improve text quality.
- **Feature Engineering** – Using TF-IDF vectorization to extract meaningful features from the text, enabling better classification.
- **Sentiment Classification** – Implementing both VADER and an ML-based approach to compare performance and tendencies in sentiment assignment.
- **Exploratory Data Analysis (EDA)** – Visualizing results through bar charts, pie charts, and heatmaps to identify patterns in sentiment distribution.
- **Subjectivity Analysis** – Differentiating between subjective and objective reviews to understand how emotions correlate with personal opinions versus factual statements.

Conclusion

The comparison between VADER and the ML model reveals distinct differences in sentiment classification. VADER tends to classify more reviews under extreme emotions like excitement and anger, whereas the ML model provides a more neutral and evenly distributed classification. Additionally, the subjectivity analysis indicates that excitement and anger are more commonly found in subjective reviews, while neutral sentiments are more balanced between objectivity and subjectivity. These findings suggest that while VADER is useful for quick polarity detection, an ML model may offer a more nuanced and context-aware approach to sentiment analysis