IMDB Movie Reviews: Comprehensive NER & POS Distribution Analysis

Abstract

This report analyzes Named Entity Recognition (NER) and Part-of-Speech (POS) distributions in 50,000 IMDB movie reviews to identify linguistic differences between positive and negative sentiments. The analysis reveals notable variations in entity usage and grammatical structures between sentiment classes.

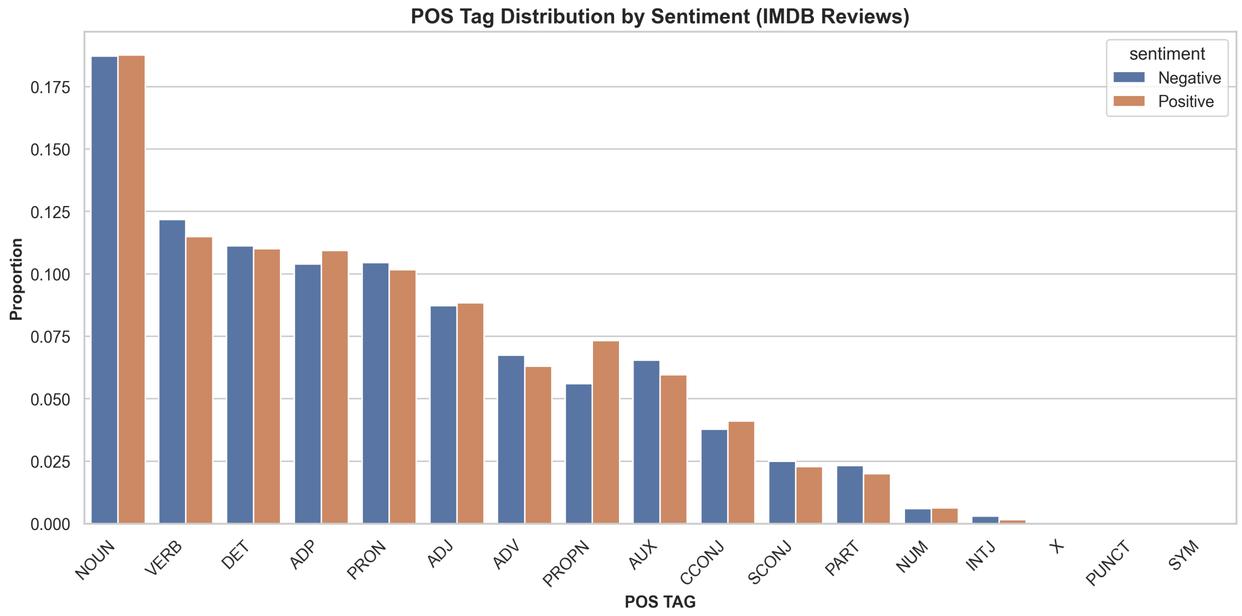
Dataset Overview

Metric	Negative Reviews	Positive Reviews
Documents	25,000	25,000
Total Tokens	5,665,558	5,752,715
Total Entities	272,404	340,782
Unique POS Tags	17	17
Unique Entity Types	18	18

Part-of-Speech Distribution Analysis

The POS distribution reveals notable differences in grammatical structure between positive and negative reviews. The following table shows the top 10 POS tags:

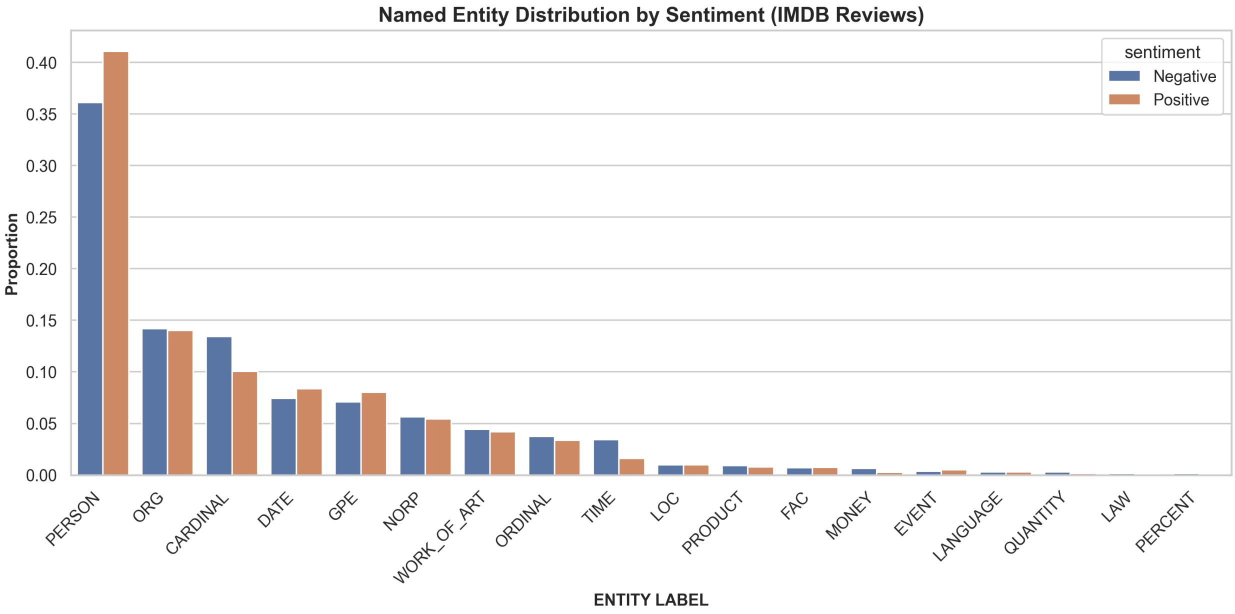
POS Tag	Negative (%)	Positive (%)	Difference
NOUN	18.73	18.77	+0.04%
VERB	12.17	11.49	-0.68%
DET	11.12	11.01	-0.11%
ADP	10.40	10.94	+0.55%
PRON	10.45	10.17	-0.28%
ADJ	8.72	8.84	+0.11%
ADV	6.74	6.30	-0.44%
PROPN	5.60	7.33	+1.74%
AUX	6.54	5.97	-0.58%
CCONJ	3.78	4.10	+0.32%



Named Entity Recognition Distribution Analysis

Named entity analysis reveals substantial differences in how positive and negative reviewers reference people, places, and other entities:

Entity Type	Negative (%)	Positive (%)	Difference
PERSON	36.10	41.05	+4.95%
ORG	14.17	14.01	-0.16%
CARDINAL	13.43	10.04	-3.39%
DATE	7.44	8.36	+0.92%
GPE	7.10	8.00	+0.90%
NORP	5.64	5.43	-0.21%
WORK_OF_ART	4.45	4.18	-0.26%
ORDINAL	3.74	3.35	-0.39%
TIME	3.43	1.60	-1.82%
LOC	0.98	0.99	+0.01%


Named Entity Distribution by Sentiment (IMDB Reviews)
sentiment
Negative
Positive
Proportion
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
PERSON
ORG
CARDINAL
DATE
GPE
NORP
WORK_OF_ART
ORDINAL
TIME
LOC
PRODUCT
FAC
MONEY
EVENT
LANGUAGE
QUANTITY
LAW
PERCENT
ENTITY LABEL

Key Findings and Patterns

Notable Differences Between Sentiments:

Positive Reviews:

- Higher entity density: 5.92% vs 4.81% (23% increase)
- More proper nouns (PROPN): 7.33% vs 5.60% (31% increase)
- More person entity mentions: 41.05% vs 36.10% (14% increase)
- Interpretation: Referential writing style that credits specific individuals

Negative Reviews:

- Higher verb usage: 12.17% vs 11.49%
- More temporal expressions: 3.43% vs 1.60%
- More cardinal numbers: 13.43% vs 10.04%
- Interpretation: Analytical approach with precise, quantitative critiques

Hypotheses and Explanations

1. Attribution Hypothesis: Positive reviewers credit specific individuals (actors, directors) for a film's success, while negative reviewers focus on abstract failures rather than blaming individuals. This explains the substantial increase in PERSON entities and proper nouns in positive reviews.

2. Specificity in Criticism: Negative reviewers employ temporal and numerical precision to support their critiques objectively. References to specific durations, scenes, or quantifiable elements provide concrete evidence for negative assessments.

3. Descriptive vs. Analytical Language: Satisfied viewers use descriptive language emphasizing what exists (nouns, adjectives) and who created it. Dissatisfied viewers use analytical language focusing on what happens (verbs) and when things go wrong (temporal expressions).

4. Emotional Expression: The higher rate of interjections in negative reviews suggests that disappointment elicits more spontaneous emotional reactions than satisfaction, which manifests in more structured, referential praise.

Conclusion

This analysis reveals notable differences in the distribution of named entities and parts of speech between positive and negative movie reviews. Positive reviews show 31% higher proper noun usage, 23% greater entity density, and 14% more person mentions, indicating a referential writing style that credits specific individuals. Negative reviews exhibit higher verb usage, more temporal expressions, and elevated cardinal numbers, reflecting an analytical approach with precise, quantitative critiques. These patterns suggest that sentiment is deeply embedded in grammatical structure and referencing behavior, not just evaluative vocabulary.