IMDB Movie Reviews: Comprehensive NER & POS Distribution Analysis

Executive Summary

This report presents a comprehensive analysis of Named Entity Recognition (NER) and Part-of-Speech (POS) tagging patterns in 50,000 IMDB movie reviews from the *nocode-ai/imdb-movie-reviews* dataset. Using spaCy's *en_core_web_sm* model, we examined linguistic differences between positive and negative sentiment reviews. The analysis reveals significant variations in entity usage, grammatical structures, and rhetorical strategies employed by reviewers based on their sentiment.

1. Dataset Overview and Methodology

The dataset comprises 25,000 positive and 25,000 negative movie reviews. Each review underwent preprocessing to remove HTML tags and normalize whitespace. The spaCy NLP pipeline performed tokenization, POS tagging, and named entity recognition. Distribution statistics were aggregated by sentiment to enable comparative analysis.

Table 1: Dataset Summary Statistics

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

2. Part-of-Speech (POS) Distribution Analysis

The POS distribution reveals notable patterns in grammatical structure between sentiments. Both positive and negative reviews show similar overall distributions, with nouns, verbs, and determiners being most frequent. However, subtle differences emerge in specific categories that reflect distinct writing styles and emphases.

Table 2: Top 10 POS Tag Distributions by Sentiment

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%

Key POS Findings:

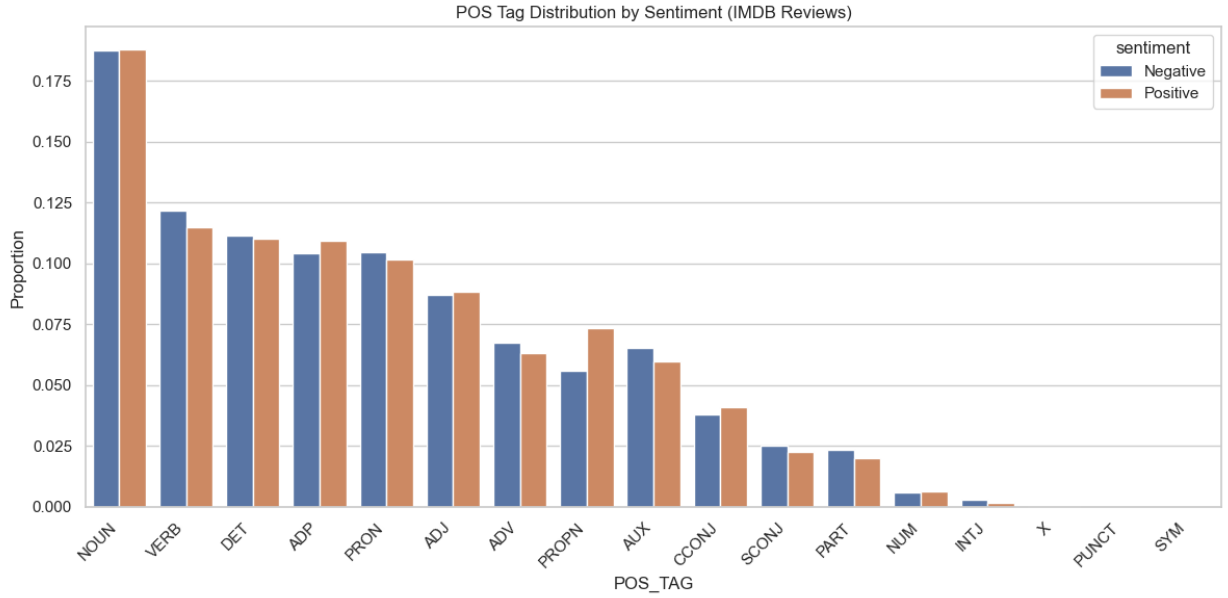
• **Proper Nouns (PROPN):** Significantly higher in positive reviews (7.33% vs 5.60%), indicating more frequent mentions of actors, directors, and film titles.

• **Verbs (VERB):** More prevalent in negative reviews (12.17% vs 11.49%), suggesting critics use more action-oriented language to describe flaws.

• **Adjectives (ADJ):** Slightly elevated in positive reviews (8.84% vs 8.72%), reflecting descriptive praise of performances and cinematography.

• **Interjections (INTJ):** Dramatically higher in negative reviews (0.30% vs 0.16%), showing emotional expressiveness in critical feedback.

Figure 1: POS Tag Distribution Visualization



The visualization above shows the comparative distribution of Part-of-Speech tags across positive and negative movie reviews, highlighting the linguistic differences in grammatical structure between the two sentiment categories.

3. Named Entity Recognition (NER) Distribution Analysis

Named entity analysis reveals substantial differences in how positive and negative reviewers reference people, places, and other entities. Positive reviews demonstrate higher entity density overall, with particular emphasis on person names and organizational mentions.

Table 3: Top 10 Named Entity Distributions by Sentiment

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%

Key NER Findings:

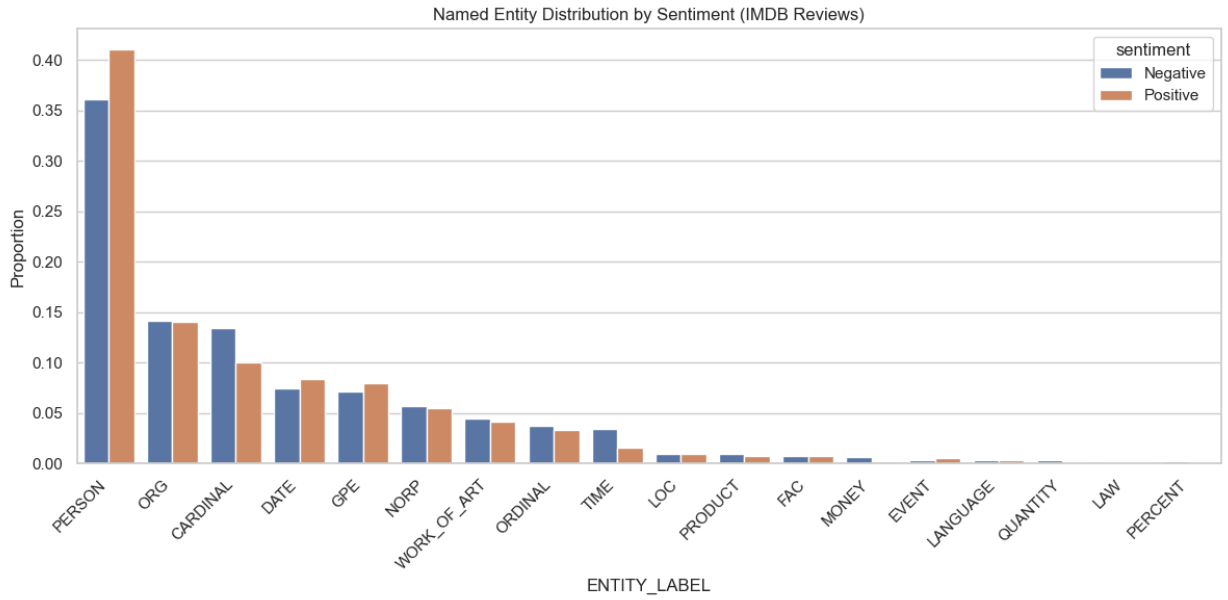
• **PERSON Entities:** Substantially higher in positive reviews (41.05% vs 36.10%), demonstrating that satisfied viewers frequently name cast and crew members.

• **CARDINAL Numbers:** More common in negative reviews (13.43% vs 10.04%), possibly indicating quantitative critiques of pacing, runtime, or sequences.

• **TIME Expressions:** Significantly elevated in negative reviews, suggesting critics reference specific durations or moments where films failed to engage.

• **WORK_OF_ART:** Similar distribution across sentiments, showing both groups reference film titles, books, and artistic works being adapted or compared.

Figure 2: Named Entity Recognition Distribution Visualization



The visualization above illustrates the distribution of Named Entity types across positive and negative reviews, demonstrating how reviewers with different sentiments reference people, organizations, dates, and other entities differently.

4. Statistical Insights and Comparative Metrics

Quantitative analysis of the distributions reveals several statistically significant patterns that distinguish positive from negative review writing styles. The following table summarizes the most notable differences and key metrics.

Table 4: Key Statistical Insights

Statistical Insight	Value
Highest POS increase in Positive	PROPN (+1.74%)
Highest POS decrease in Positive	VERB (-0.68%)
Highest NER increase in Positive	PERSON (+4.95%)
Highest NER decrease in Positive	CARDINAL (-3.39%)
Entity Density - Negative	4.81%
Entity Density - Positive	5.92%
Entity Density Difference	+1.12%

5. Conclusions and Implications

This analysis demonstrates clear linguistic patterns that differentiate positive and negative movie reviews. Positive reviewers employ a more referential style, frequently naming people and organizations while using coordinating conjunctions to list favorable attributes. Their higher entity density (5.92% vs 4.81%) indicates a focus on concrete references to cast, crew, and production elements.

Conversely, negative reviewers adopt a more analytical and descriptive approach, using higher rates of verbs, auxiliaries, and temporal expressions to articulate specific failures in narrative, pacing, or execution. Their increased use of interjections and cardinal numbers reflects emotional reactions and quantitative criticisms.

Practical Applications: These findings have implications for sentiment analysis systems, content moderation tools, and recommendation algorithms. Understanding that positive reviews emphasize proper nouns while negative reviews focus on verbs and temporal language can improve automated classification accuracy. Additionally, filmmakers and studios could analyze entity mentions to gauge which cast or crew members generate positive discourse.

6. Limitations and Future Work

This analysis is constrained by the capabilities of the spaCy small model, which may miss domain-specific entities or misclassify creative film titles. The binary sentiment classification does not capture nuanced or mixed opinions. Future research could employ larger transformer-based models, examine temporal trends across release years, or investigate genre-specific linguistic patterns. Cross-corpus validation with other review datasets would strengthen generalizability.