Dataset Analysis Report

Dataset Path: binary_class_dataset

Dataset Overview:

• Total samples: 7,908
• Total entities: 308,853
• Number of splits: 3
• Unique labels: 2

Split Details:

• train: 6,326 samples
• validation: 790 samples
• test: 792 samples

Label Information:

• O: 227,079 occurrences
• sensitive_data: 81,774 occurrences

Split Statistics

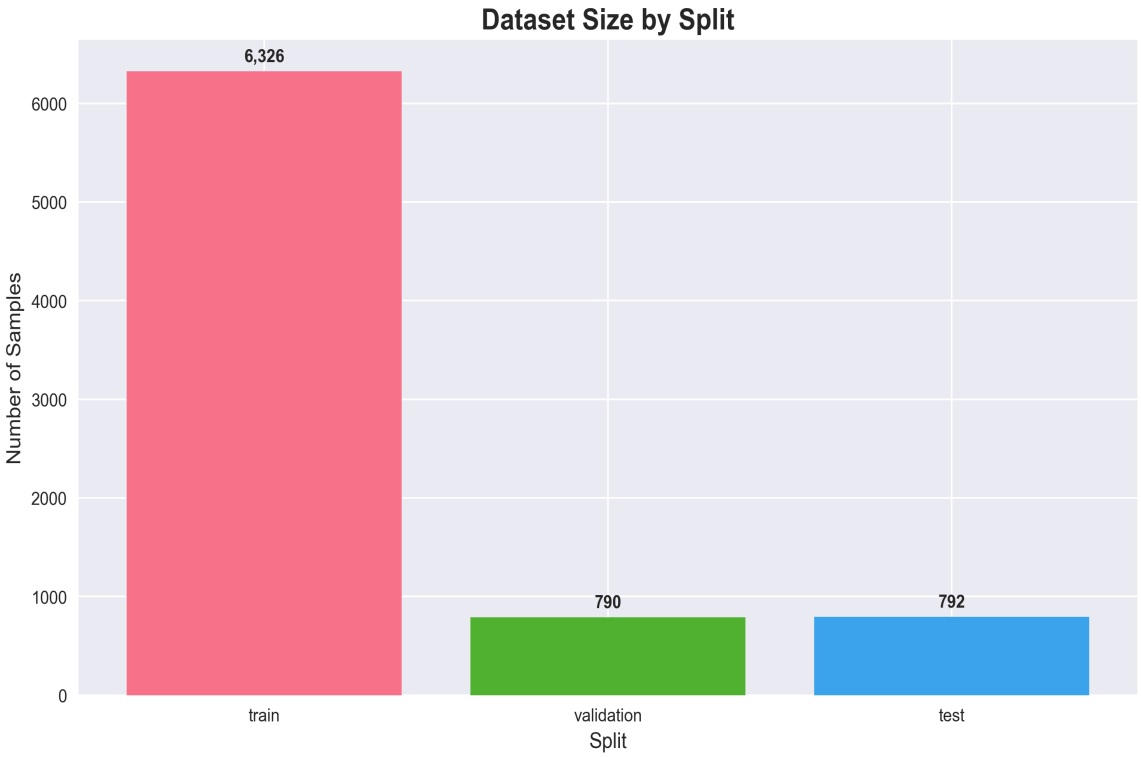
Split	Samples	Entities
train	6,326	247,118
validation	790	31,105
test	792	30,630

Label Distribution by Split

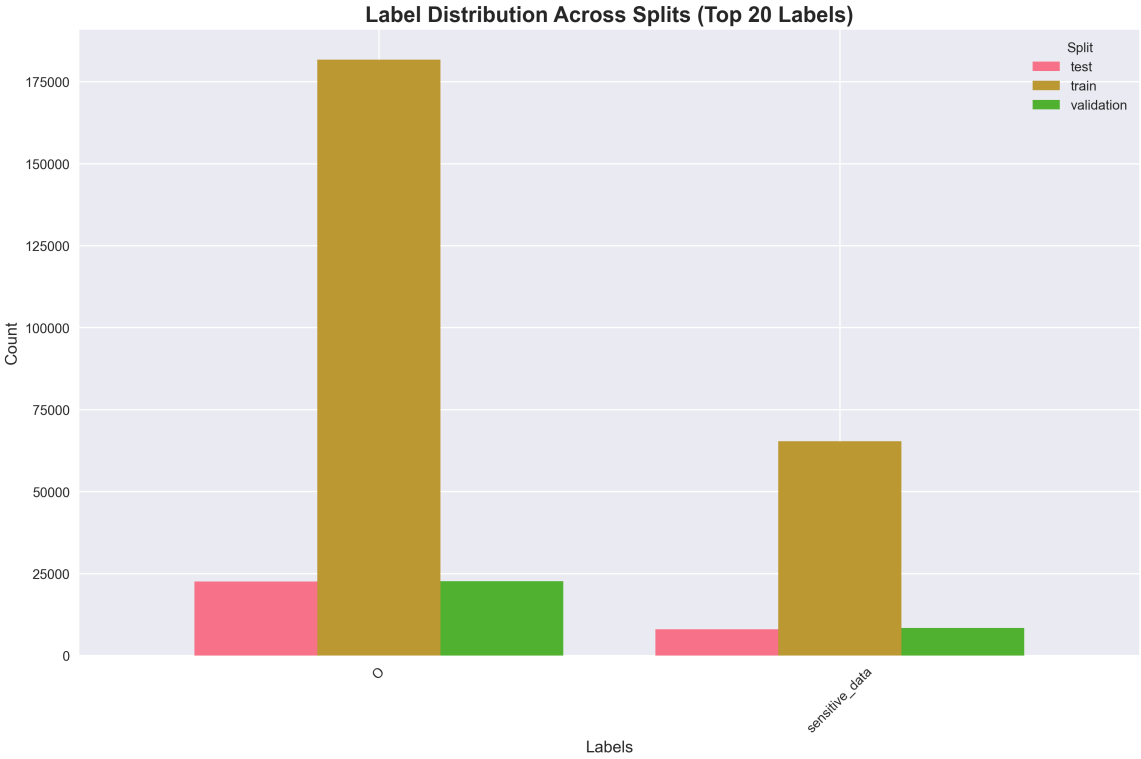
Label	train	validation	test
O	181,738	22,714	22,627
sensitive_data	65,380	8,391	8,003

Visualizations

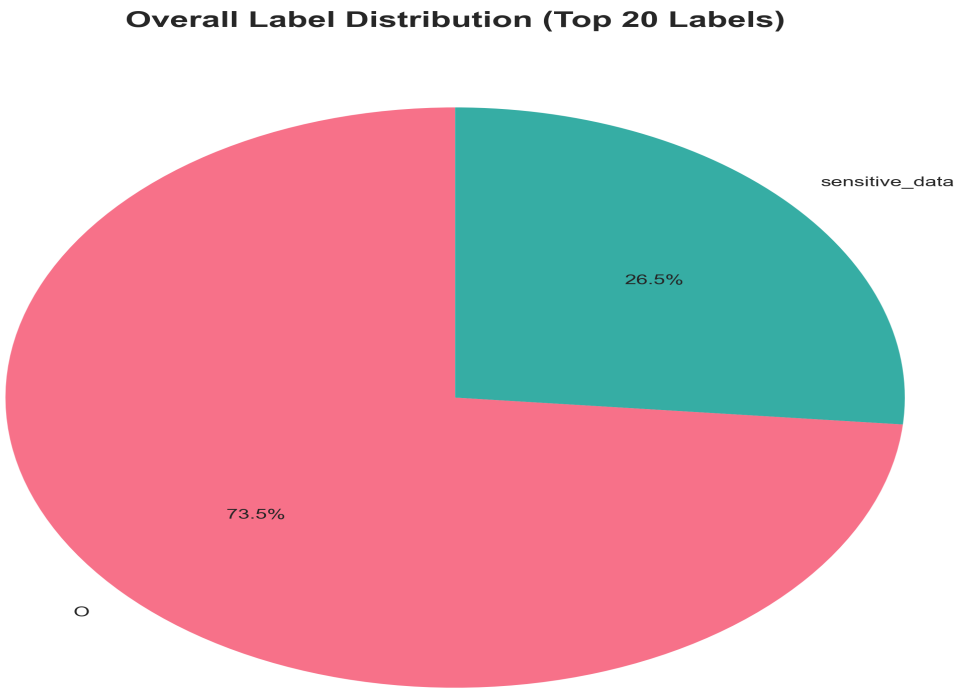
Dataset Size Comparison by Split



Label Distribution Across Splits (Top 20 Labels)



Overall Label Distribution (Top 20 Labels)



Total Label Frequencies (Top 20 Labels)