Dataset Analysis Report

Dataset Path: ner_dataset_80_10_10

Dataset Overview:

• Total samples: 7,908
• Total entities: 308,853
• Number of splits: 3
• Unique labels: 26

Split Details:

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

Label Information:

• ADDRESS: 15,826 occurrences
• BANK_ACCOUNT: 2,443 occurrences
• BIOMETRIC_DATA: 36 occurrences
• BIRTH_DATE_PLACE: 8,379 occurrences
• CREDIT_SCORE: 40 occurrences
• CRIMINAL_RECORD: 7 occurrences
• EMAIL: 522 occurrences
• FINANCIAL_DATA: 5,622 occurrences
• GENDER: 252 occurrences
• HEALTH_INFO: 1,272 occurrences
• IBAN: 54 occurrences
• INCOME_EXPENSE: 2,235 occurrences
• IP_ADDRESS: 32 occurrences
• LOCATION: 646 occurrences
• NAME: 14,083 occurrences
• O: 227,079 occurrences
• PHONE: 11,938 occurrences
• POLITICAL_OPINION: 4 occurrences
• RACE: 16 occurrences
• RELIGION_OR_SECT: 163 occurrences
• SEXUAL_LIFE: 88 occurrences
• SOCIAL_MEDIA_ACCOUNT: 2,741 occurrences
• SURNAME: 11,061 occurrences
• TCKN: 4,043 occurrences
• TRANSACTION_HISTORY: 235 occurrences
• UNION_MEMBERSHIP: 36 occurrences

Split Statistics

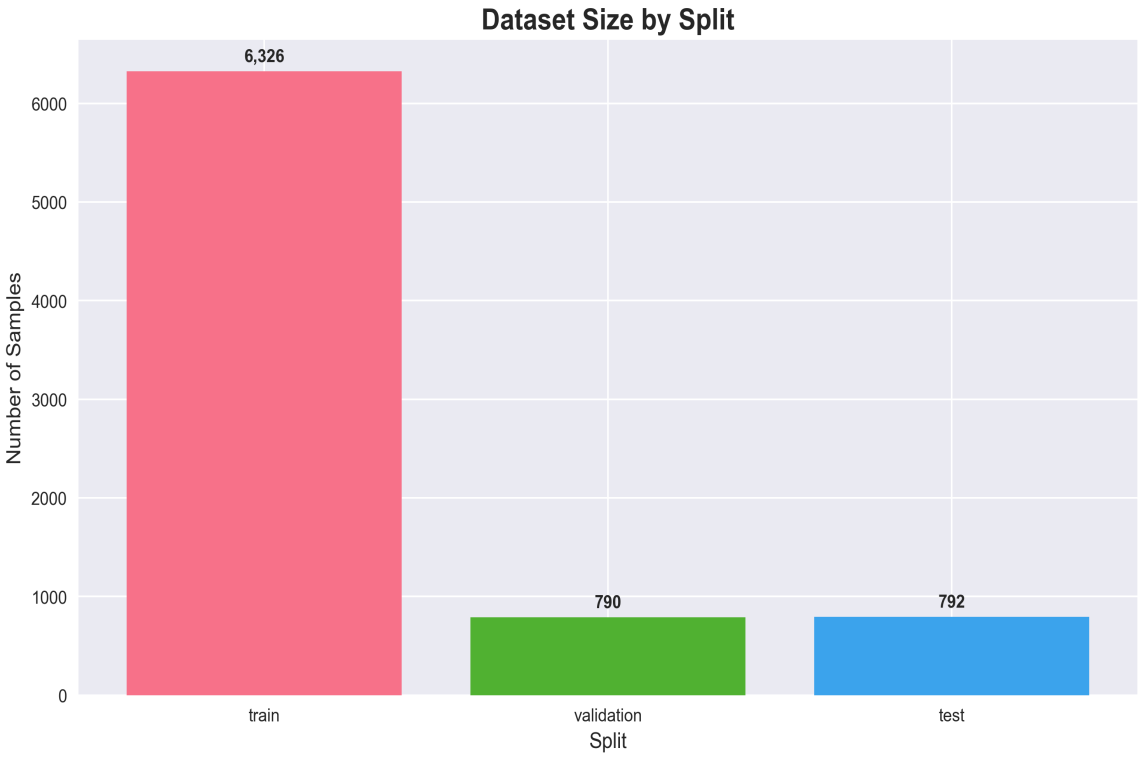
Split	Samples	Entities
train	6,326	247,118
validation	790	31,105
test	792	30,630

Label Distribution by Split

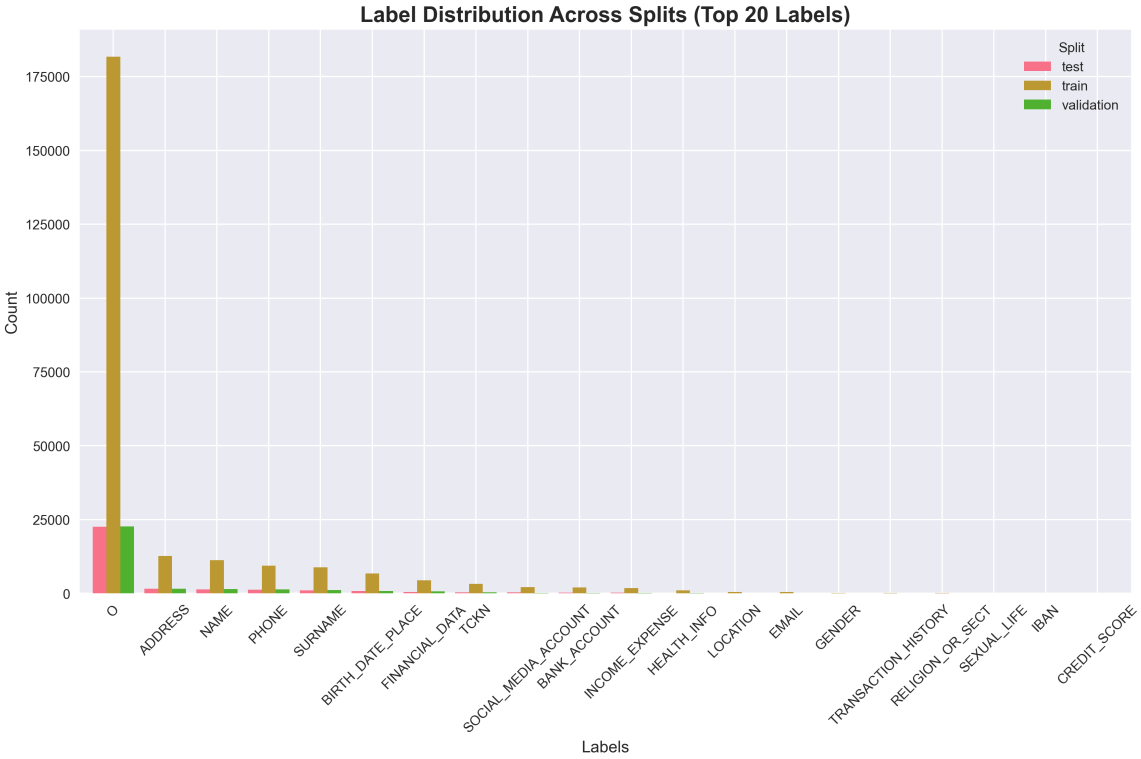
Label	train	validation	test
ADDRESS	12,684	1,577	1,565
BANK_ACCOUNT	2,049	176	218
BIOMETRIC_DATA	30	4	2
BIRTH_DATE_PLACE	6,775	837	767
CREDIT_SCORE	28	0	12
CRIMINAL_RECORD	5	2	0
EMAIL	437	42	43
FINANCIAL_DATA	4,481	671	470
GENDER	207	22	23
HEALTH_INFO	981	200	91
IBAN	30	11	13
INCOME_EXPENSE	1,773	205	257
IP_ADDRESS	8	8	16
LOCATION	526	55	65
NAME	11,259	1,441	1,383
O	181,738	22,714	22,627
PHONE	9,404	1,328	1,206
POLITICAL_OPINION	4	0	0
RACE	16	0	0
RELIGION_OR_SECT	131	15	17
SEXUAL_LIFE	66	10	12
SOCIAL_MEDIA_ACCOUNT	2,178	196	367
SURNAME	8,858	1,150	1,053
TCKN	3,252	415	376
TRANSACTION_HISTORY	180	26	29
UNION_MEMBERSHIP	18	0	18

Visualizations

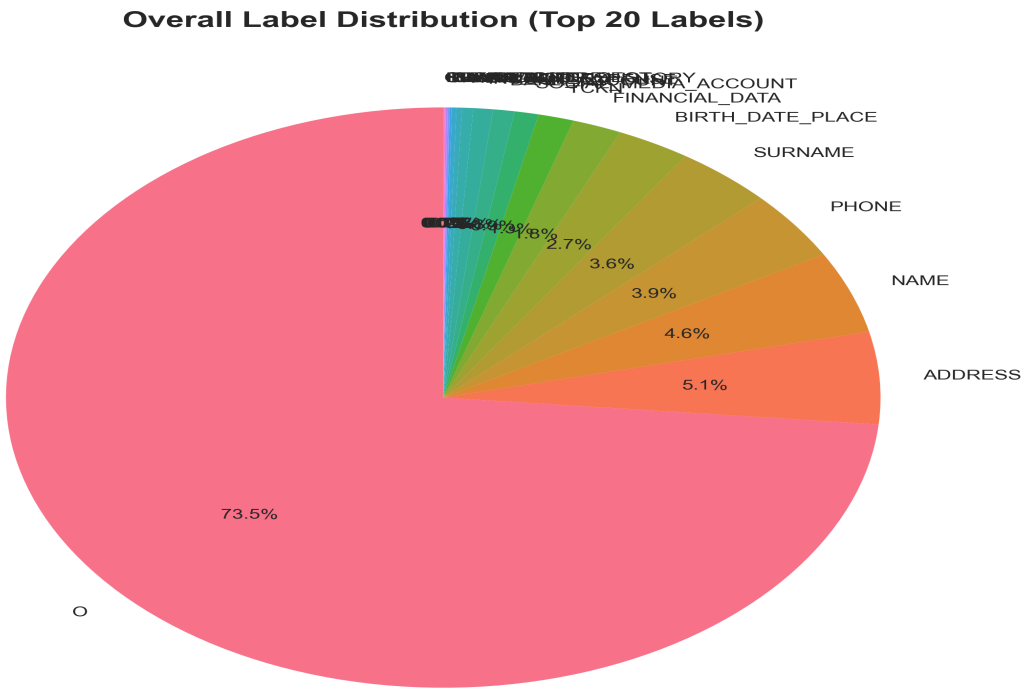
Dataset Size Comparison by Split



Label Distribution Across Splits (Top 20 Labels)



Overall Label Distribution (Top 20 Labels)



Total Label Frequencies (Top 20 Labels)