Dataset Analysis Report

Dataset Path: merged_ner_dataset

Dataset Overview:

• Total samples: 9,891
• Total entities: 372,878
• Number of splits: 3
• Unique labels: 28

Split Details:

• train: 7,913 samples
• validation: 989 samples
• test: 989 samples

Label Information:

• ADDRESS: 17,586 occurrences
• BANK_ACCOUNT: 2,225 occurrences
• BIOMETRIC_DATA: 42 occurrences
• BIRTH_DATE_PLACE: 9,800 occurrences
• CREDIT_CARD_NUMBER: 2 occurrences
• CREDIT_SCORE: 28 occurrences
• CRIMINAL_RECORD: 24 occurrences
• EMAIL: 959 occurrences
• ETHNIC_ORIGIN: 207 occurrences
• FINANCIAL_DATA: 5,242 occurrences
• GENDER: 416 occurrences
• HEALTH_INFO: 1,246 occurrences
• IBAN: 46 occurrences
• INCOME_EXPENSE: 2,016 occurrences
• IP_ADDRESS: 16 occurrences
• LOCATION: 1,190 occurrences
• NAME: 16,701 occurrences
• O: 276,055 occurrences
• PHONE: 19,401 occurrences
• POLITICAL_OPINION: 18 occurrences
• RACE: 16 occurrences
• RELIGION_OR_SECT: 146 occurrences
• SEXUAL_LIFE: 79 occurrences
• SOCIAL_MEDIA_ACCOUNT: 2,724 occurrences
• SURNAME: 12,634 occurrences
• TCKN: 3,706 occurrences
• TRANSACTION_HISTORY: 206 occurrences
• UNION_MEMBERSHIP: 147 occurrences

Split Statistics

Split	Samples	Entities
train	7,913	297,871
validation	989	37,586

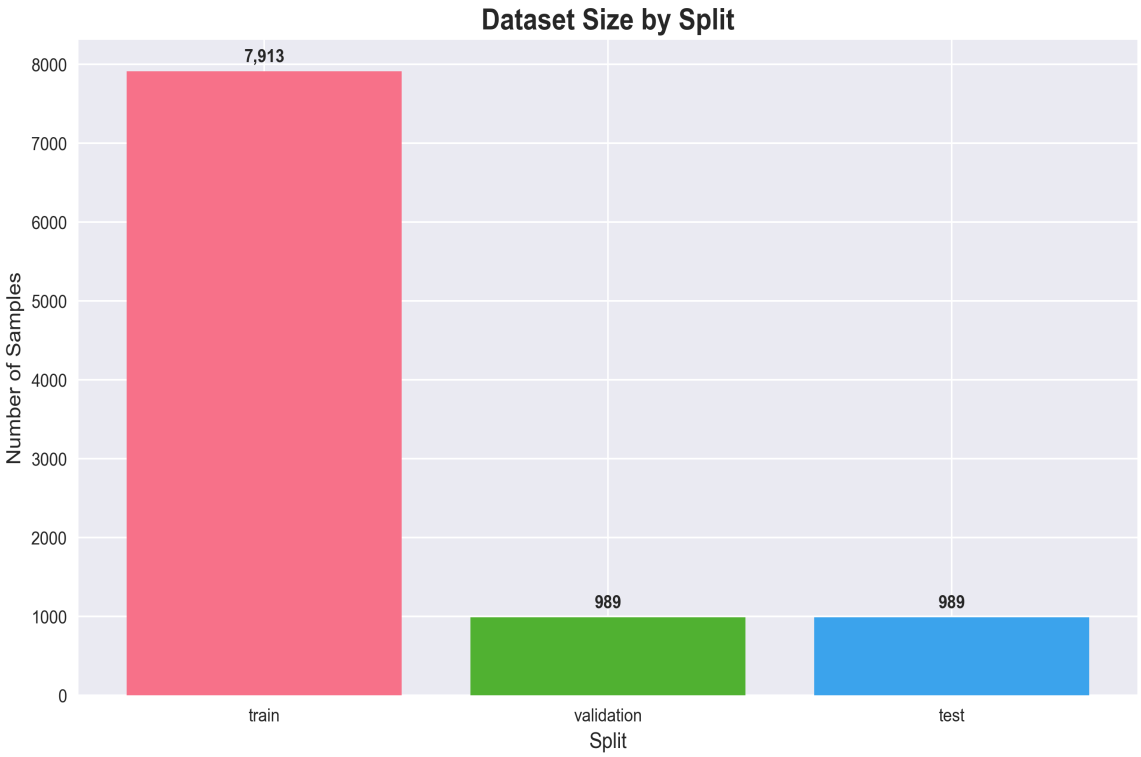
test	989	37,421

Label Distribution by Split

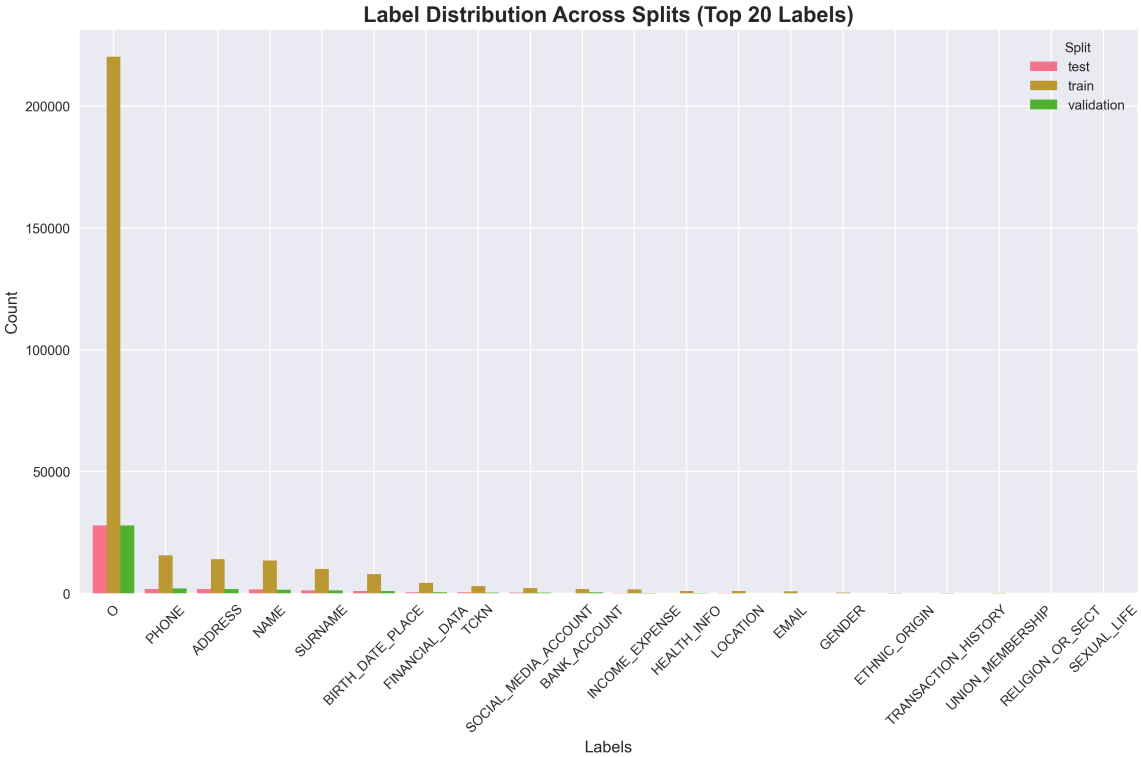
Label	train	validation	test
ADDRESS	14,039	1,797	1,750
BANK_ACCOUNT	1,720	391	114
BIOMETRIC_DATA	36	2	4
BIRTH_DATE_PLACE	7,928	941	931
CREDIT_CARD_NUMBER	2	0	0
CREDIT_SCORE	20	8	0
CRIMINAL_RECORD	18	6	0
EMAIL	796	101	62
ETHNIC_ORIGIN	178	18	11
FINANCIAL_DATA	4,300	477	465
GENDER	337	41	38
HEALTH_INFO	987	140	119
IBAN	35	5	6
INCOME_EXPENSE	1,630	172	214
IP_ADDRESS	8	0	8
LOCATION	947	107	136
NAME	13,440	1,584	1,677
O	220,236	27,891	27,928
PHONE	15,579	2,000	1,822
POLITICAL_OPINION	12	4	2
RACE	10	2	4
RELIGION_OR_SECT	117	7	22
SEXUAL_LIFE	62	5	12
SOCIAL_MEDIA_ACCOUNT	2,123	279	322
SURNAME	10,083	1,251	1,300
TCKN	2,935	330	441
TRANSACTION_HISTORY	151	27	28
UNION_MEMBERSHIP	142	0	5

Visualizations

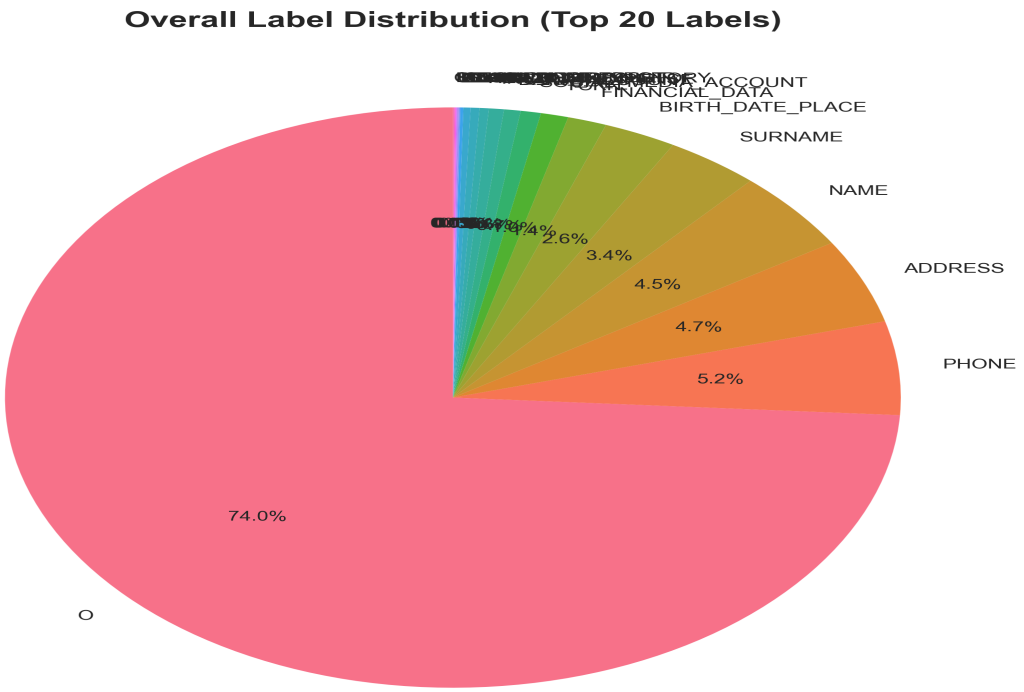
Dataset Size Comparison by Split



Label Distribution Across Splits (Top 20 Labels)



Overall Label Distribution (Top 20 Labels)



Total Label Frequencies (Top 20 Labels)