Dataset Analysis Report

Dataset Path: merged_ner_dataset

Dataset Overview:

• Total samples: 7,123
• Total entities: 278,299
• Number of splits: 3
• Unique labels: 26

Split Details:

• train: 5,699 samples
• validation: 712 samples
• test: 712 samples

Label Information:

• ADDRESS: 14,261 occurrences
• BANK_ACCOUNT: 2,225 occurrences
• BIOMETRIC_DATA: 34 occurrences
• BIRTH_DATE_PLACE: 7,618 occurrences
• CREDIT_SCORE: 28 occurrences
• CRIMINAL_RECORD: 7 occurrences
• EMAIL: 479 occurrences
• FINANCIAL_DATA: 5,152 occurrences
• GENDER: 229 occurrences
• HEALTH_INFO: 1,181 occurrences
• IBAN: 41 occurrences
• INCOME_EXPENSE: 1,978 occurrences
• IP_ADDRESS: 16 occurrences
• LOCATION: 581 occurrences
• NAME: 12,708 occurrences
• O: 204,514 occurrences
• PHONE: 10,732 occurrences
• POLITICAL_OPINION: 4 occurrences
• RACE: 16 occurrences
• RELIGION_OR_SECT: 146 occurrences
• SEXUAL_LIFE: 76 occurrences
• SOCIAL_MEDIA_ACCOUNT: 2,374 occurrences
• SURNAME: 10,008 occurrences
• TCKN: 3,667 occurrences
• TRANSACTION_HISTORY: 206 occurrences
• UNION_MEMBERSHIP: 18 occurrences

Split Statistics

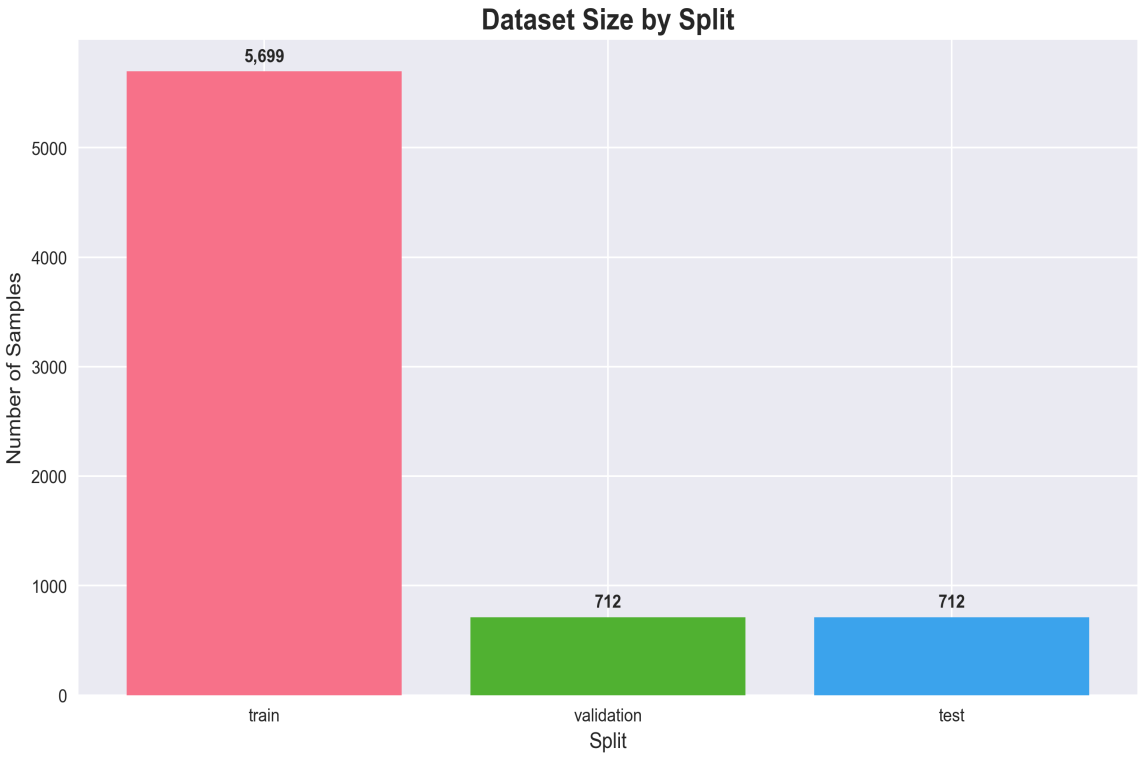
Split	Samples	Entities
train	5,699	222,775
validation	712	27,965
test	712	27,559

Label Distribution by Split

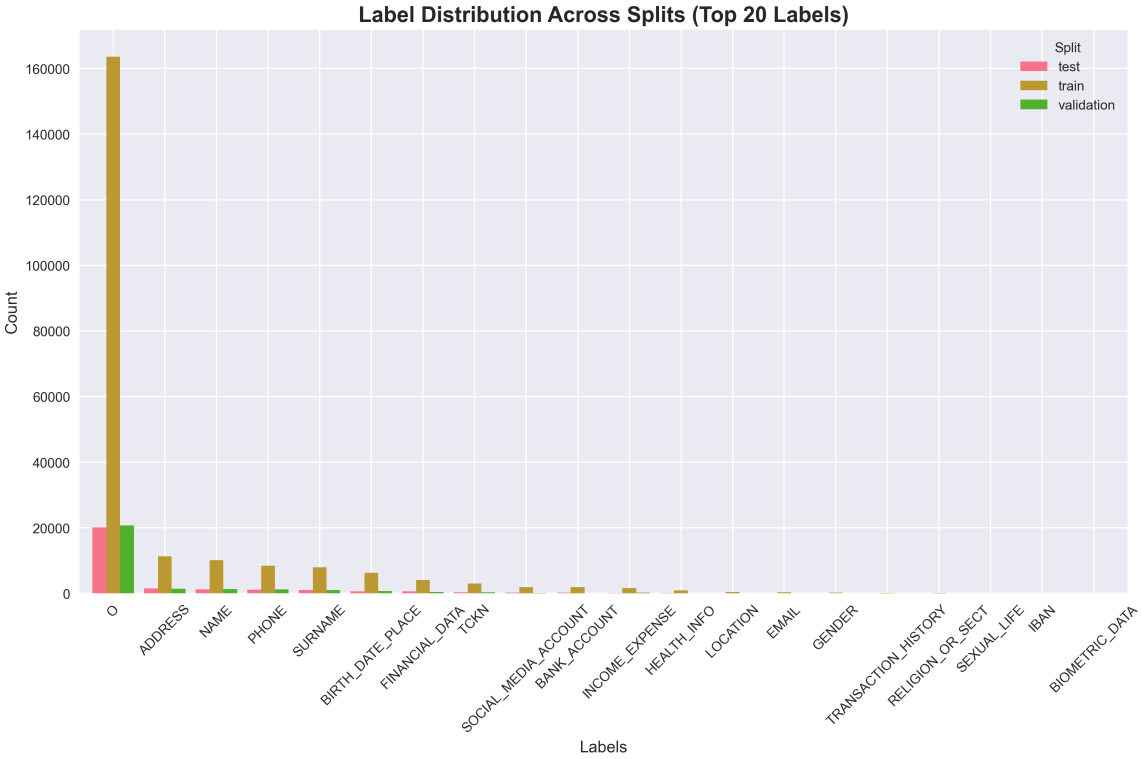
Label	train	validation	test
ADDRESS	11,328	1,446	1,487
BANK_ACCOUNT	1,899	83	243
BIOMETRIC_DATA	24	4	6
BIRTH_DATE_PLACE	6,266	695	657
CREDIT_SCORE	24	4	0
CRIMINAL_RECORD	4	3	0
EMAIL	383	64	32
FINANCIAL_DATA	4,066	467	619
GENDER	190	23	16
HEALTH_INFO	947	73	161
IBAN	18	12	11
INCOME_EXPENSE	1,620	192	166
IP_ADDRESS	16	0	0
LOCATION	440	79	62
NAME	10,162	1,339	1,207
O	163,613	20,728	20,173
PHONE	8,427	1,194	1,111
POLITICAL_OPINION	4	0	0
RACE	11	5	0
RELIGION_OR_SECT	124	10	12
SEXUAL_LIFE	66	7	3
SOCIAL_MEDIA_ACCOUNT	1,931	173	270
SURNAME	7,967	1,039	1,002
TCKN	3,050	311	306
TRANSACTION_HISTORY	177	14	15
UNION_MEMBERSHIP	18	0	0

Visualizations

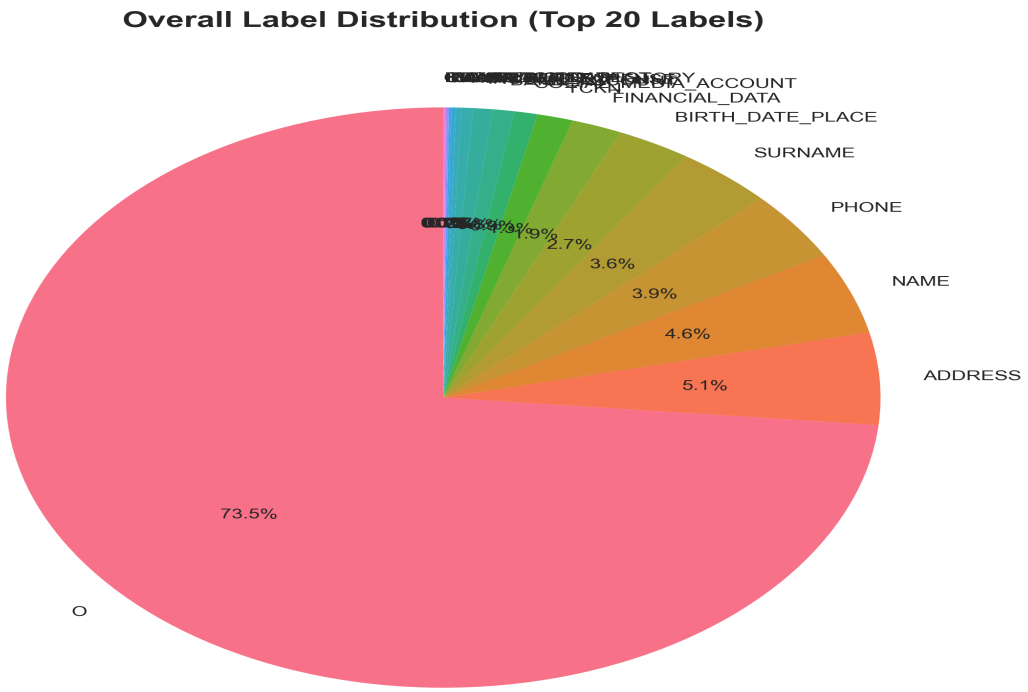
Dataset Size Comparison by Split



Label Distribution Across Splits (Top 20 Labels)



Overall Label Distribution (Top 20 Labels)



Total Label Frequencies (Top 20 Labels)