Dataset Analysis Report

Dataset Path: outputs\hf_ner_dataset_exact_match

Dataset Overview:

• Total samples: 2,775
• Total entities: 94,655
• Number of splits: 2
• Unique labels: 22

Split Details:

• train: 2,497 samples
• validation: 278 samples

Label Information:

• O: 71,603 occurrences
• address: 3,325 occurrences
• biometric_data: 8 occurrences
• birth_date_place: 2,188 occurrences
• credit_card_number: 2 occurrences
• criminal_record: 17 occurrences
• email: 480 occurrences
• ethnic_origin: 207 occurrences
• financial_data: 90 occurrences
• gender: 187 occurrences
• health_info: 65 occurrences
• iban: 5 occurrences
• income_expense: 38 occurrences
• location: 609 occurrences
• name: 4,001 occurrences
• phone: 8,669 occurrences
• political_opinion: 14 occurrences
• sexual_life: 3 occurrences
• social_media_account: 350 occurrences
• surname: 2,626 occurrences
• tckn: 39 occurrences
• union_membership: 129 occurrences

Split Statistics

Split	Samples	Entities
train	2,497	85,354
validation	278	9,301

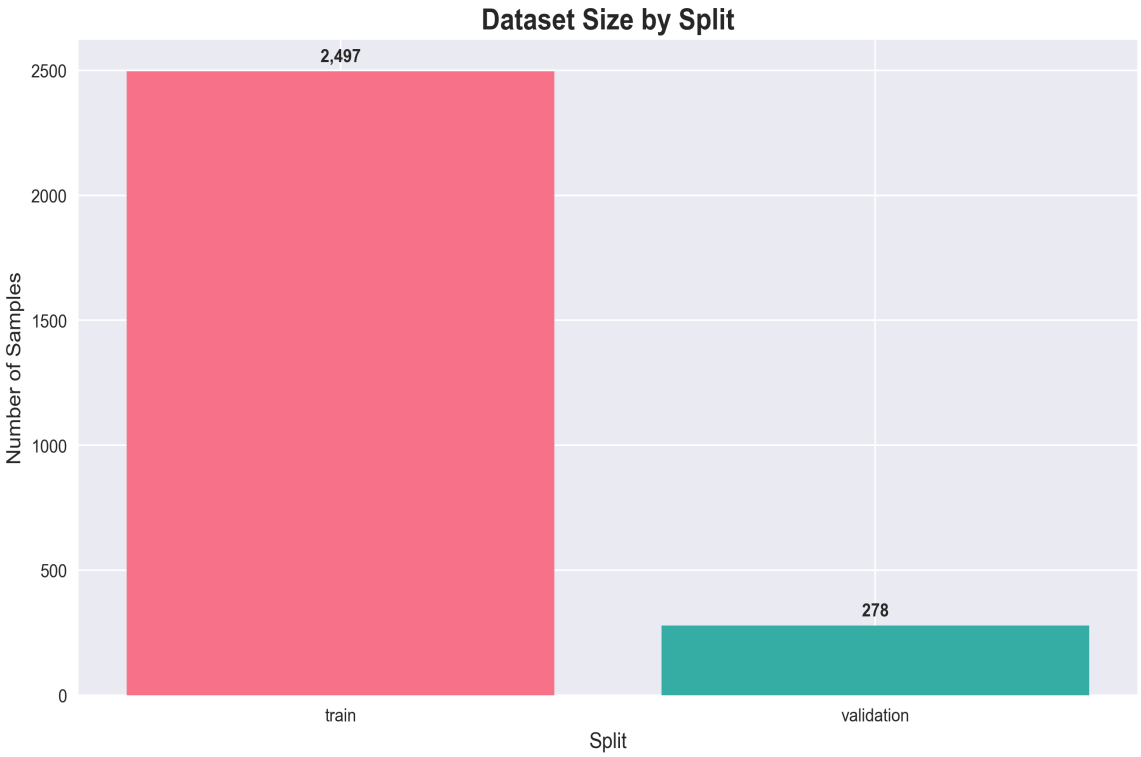
Label Distribution by Split

Label	train	validation
O	64,534	7,069

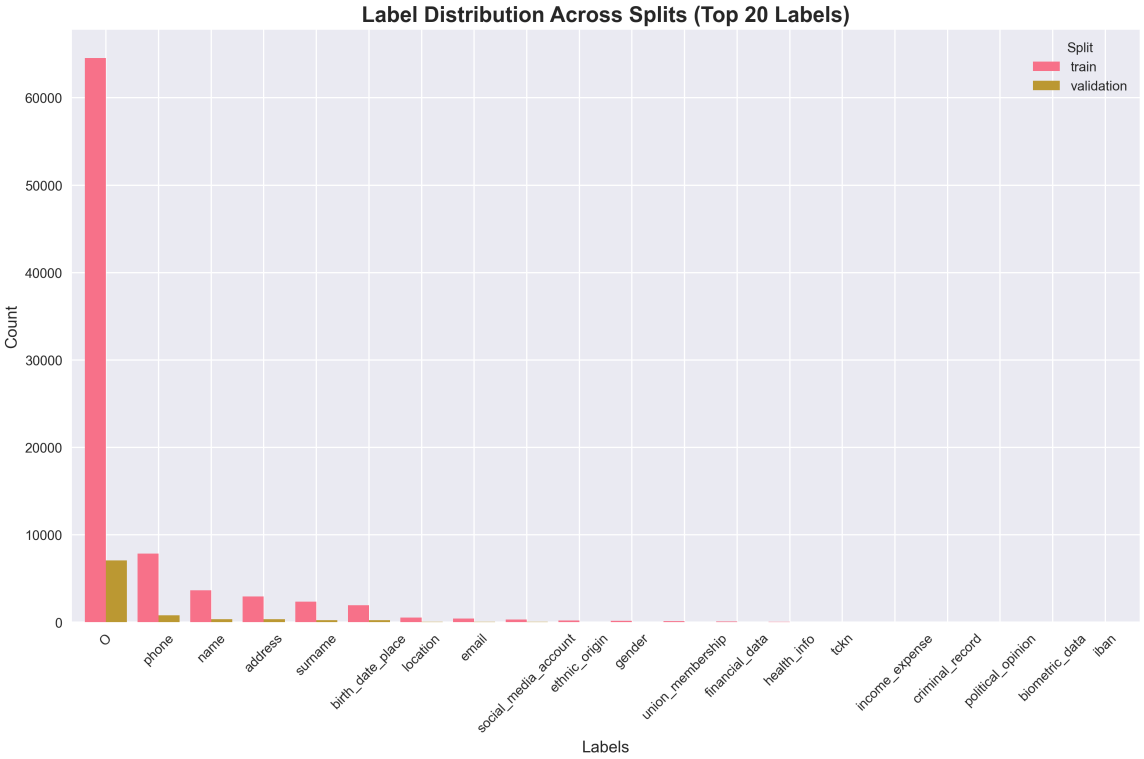
address	2,970	355
biometric_data	8	0
birth_date_place	1,952	236
credit_card_number	2	0
criminal_record	14	3
email	418	62
ethnic_origin	191	16
financial_data	86	4
gender	169	18
health_info	55	10
iban	5	0
income_expense	34	4
location	549	60
name	3,655	346
phone	7,852	817
political_opinion	14	0
sexual_life	2	1
social_media_account	305	45
surname	2,377	249
tckn	33	6
union_membership	129	0

Visualizations

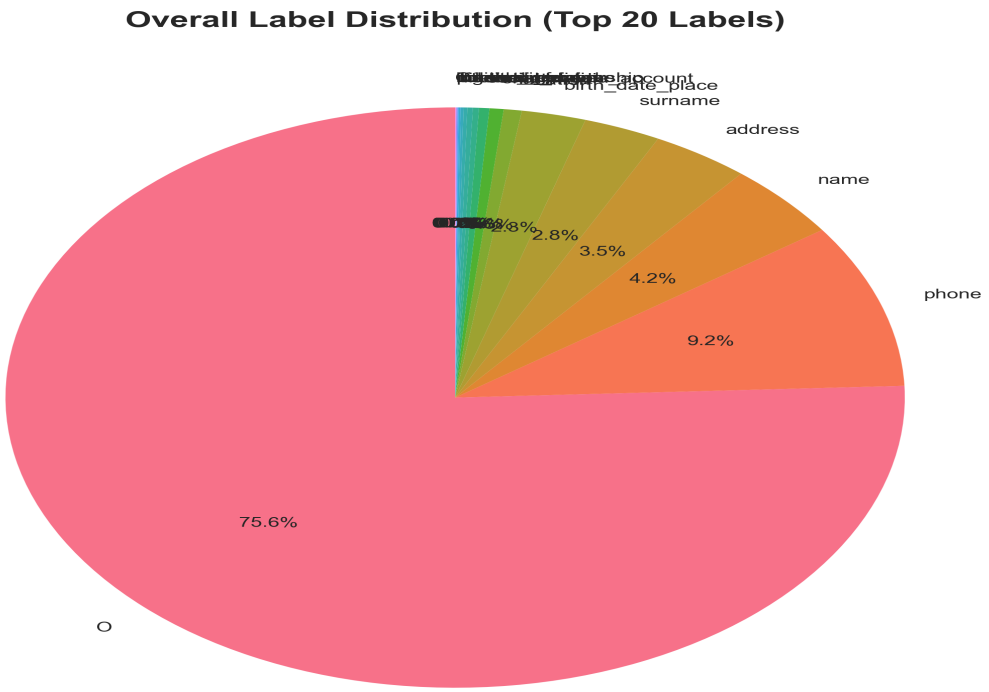
Dataset Size Comparison by Split



Label Distribution Across Splits (Top 20 Labels)



Overall Label Distribution (Top 20 Labels)



Total Label Frequencies (Top 20 Labels)