**Statistics of ProTrix Contrastive Dataset:**

**Basic Statistics:**

* Total examples: 3157
* Total positive pairs: 31570
* Total negative pairs: 30444
* Average positives per anchor: 10.00
* Average negatives per anchor: 9.64

**Distance Statistics:**

* **Positive Pairs**:
  + min: 0.000
  + max: 18.111
  + mean: 0.697
  + median: 0.484
  + std: 0.950
* **Negative Pairs (Hard + Extreme):**
  + min: 0.019
  + max: 45.363
  + mean: 18.313
  + median: 21.489
  + std: 14.783
* **Negative Distance Ranges:**
  + 0.0 - 2.5: 2372
  + 2.5 - 5.0: 12287
  + >10.0 : 15785

**Topic Distribution:**

* **Broad Topics (8):**
  + Sports (5): 855 (27.1%)
  + Music Charts: 745 (23.6%)
  + Sports Leagues: 407 (12.9%)
  + Entertainment Industry: 405 (12.8%)
  + Sports (4): 358 (11.3%)
  + College Sports: 198 (6.3%)
  + College Football (2): 102 (3.2%)
  + Telugu Cinema: 87 (2.8%)
* **Medium Topics (63):**
  + Company News: 117 (3.7%)
  + Sports-2: 100 (3.2%)
  + Politics: 91 (2.9%)
  + Actresses: 87 (2.8%)
  + Census Data: 83 (2.6%)
  + Historic House: 79 (2.5%)
  + Science Fiction: 74 (2.3%)
  + College Football: 73 (2.3%)
  + Hockey News: 71 (2.2%)
  + Music Industry: 70 (2.2%)
  + Religious History: 70 (2.2%)
  + University Finance: 68 (2.2%)
  + Biochemistry: 67 (2.1%)
  + Winter Sports: 67 (2.1%)
  + Olympic Games: 66 (2.1%)
  + Formula Racing: 64 (2.0%)
  + Lacrosse Teams: 62 (2.0%)
  + Sports News-2: 61 (1.9%)
  + Rock Music: 60 (1.9%)
  + Film Awards: 60 (1.9%)
  + Sports News-3: 57 (1.8%)
  + Sports: 57 (1.8%)
  + College Teams: 56 (1.8%)
  + Government: 53 (1.7%)
  + Aircraft Industry: 53 (1.7%)
  + Geopolitics: 53 (1.7%)
  + European Politics: 52 (1.6%)
  + Rugby: 52 (1.6%)
  + Software Development: 49 (1.6%)
  + Cycling: 49 (1.6%)
  + Sports Teams-3: 48 (1.5%)
  + NASCAR Racing: 48 (1.5%)
  + Sports News: 48 (1.5%)
  + Characters: 47 (1.5%)
  + Sports-3: 47 (1.5%)
  + Places-2: 47 (1.5%)
  + Airports: 47 (1.5%)
  + Geography: 45 (1.4%)
  + TV Shows: 45 (1.4%)
  + Places: 42 (1.3%)
  + Latin American TV: 41 (1.3%)
  + Geography-2: 41 (1.3%)
  + Anime: 40 (1.3%)
  + Rugby Challenge: 40 (1.3%)
  + Public Transport: 39 (1.2%)
  + Entertainment: 38 (1.2%)
  + Professional Wrestling: 34 (1.1%)
  + Entertainment Awards: 34 (1.1%)
  + Film Directors: 33 (1.0%)
  + Country Music: 33 (1.0%)
  + Sports Teams-2: 31 (1.0%)
  + Food: 28 (0.9%)
  + Bollywood Songs: 27 (0.9%)
  + Montenegrin: 27 (0.9%)
  + College Athletics: 27 (0.9%)
  + Cricket News: 25 (0.8%)
  + Tennis News: 22 (0.7%)
  + Sports Teams: 19 (0.6%)
  + Education: 19 (0.6%)
  + College History: 19 (0.6%)
  + Collegiate Sports: 18 (0.6%)
  + Greyhound Racing: 14 (0.4%)
  + African Colleges: 13 (0.4%)
  + Ireland Geography: 10 (0.3%)

**Cross-topic Statistics:**

* **Broad Topic Relationships:**
  + Same broad topic negatives: 0 (0.0%)
  + Different broad topic negatives: 30444 (100.0%)
* **Medium Topic Relationships:**
  + Same medium topic negatives: 0 (0.0%)
  + Different medium topic negatives: 30444 (100.0%)

**Interpretation:**

1. **Dataset Size and Balance**

* Total examples: 3,157 anchor points
* Very balanced positive-to-negative ratio:
  + 10.0 positives per anchor (31,570 total)
  + 9.64 negatives per anchor (30,444 total)

1. Distance Distribution Analysis

* Positive pairs show tight clustering:
  + Mean distance: 0.697
  + Small standard deviation: 0.950
  + Most positive pairs are close (max: 18.111)
* Negative pairs show good separation:
  + Mean distance: 18.313
  + Larger spread (std: 14.783)
  + Most negatives are in 10.0+ range (15,785 pairs)

1. **Topic Distribution:**

* Broad Topic Distribution Characteristics:
  + 8 broad topics with uneven distribution
  + Largest: "Sports (5)" with 855 examples
  + Smallest: "Telugu Cinema" with 87 examples
* Medium Topic Distribution Characteristics:
* We have 63 distinct medium topics
* Distribution is **relatively balanced** but has some variations:
  + Largest: "Company News" (117 examples, 3.7%)
  + Smallest: "Ireland Geography" (10 examples, 0.3%)
* Most topics fall between 1-3% of the dataset
* **Good spread across different domains** (sports, entertainment, news, etc.)

**Important Observations:**

* No same-topic negatives (0%) in both broad and medium topics
* All negative pairs cross topic boundaries
* Distance distribution shows good separation between positive and negative pairs
* Very balanced positive-to-negative ratio (~10 pairs per anchor)