BPS QUESTION DEDUPLICATION SYSTEM

A Semantic Similarity Approach for Cross-Survey Redundancy Detection

Author: Sigit Nugroho Putra
Statistics Indonesia (BPS) — 2025

Abstract

National statistical offices collect large volumes of survey data across multiple programs and directorates. This often leads to duplicated question items, fragmented metadata, and inconsistent indicator definitions.
This technical report presents a semantic-similarity–based system for detecting redundant survey questions across BPS survey instruments using modern NLP embeddings (SentenceTransformers), cosine similarity, kNN retrieval, and graph clustering.
The system provides a scalable foundation for survey harmonization, metadata integration, and respondent burden reduction.

1. Introduction

Statistics Indonesia (BPS) conducts numerous national surveys such as Susenas, Sakernas, Podes, SP2KP, and Sensus Ekonomi. Each survey is managed by different directorates, resulting in:

- duplicated indicators,
- inconsistent question phrasing,
- difficulty in integrating survey metadata,
- increased respondent burden.

This report proposes an automated semantic matching system to detect and cluster similar questions across surveys.

2. Related Work

2.1 Semantic Text Similarity

- Reimers & Gurevych (2019): Semantic sentence embeddings (SBERT)
- MiniLM models for multilingual settings

- Applications in document alignment and duplicate detection

2.2 Distributed Optimization (Relevance to KAUST)

- Richtárik et al.: federated, sparse, and communication-efficient optimization
- Useful for scaling the system to BPS's multi-directorate architecture

3. Methodology

3.1 Overview Pipeline

1. Import all survey question texts
2. Generate embeddings (MiniLM-L6-v2)
3. Retrieve candidate neighbors (kNN)
4. Compute cosine similarity
5. Apply threshold ≥ 0.78
6. Group results via graph clustering
7. Export similarity pairs & clusters

3.2 Embedding Model

Model: **all-MiniLM-L6-v2**
Vector size: 384
Advantages:

- fast
- strong semantic capture
- robust for Indonesian text

3.3 Similarity Scoring

Cosine similarity between embeddings.
Redundancy threshold: **0.78**

3.4 Clustering

Graph-based connected components.

4. Experiments

4.1 Dataset

50 mixed survey questions from:

- Susenas
- Sakernas
- Podes
- Susenas modules
- SP2KP
- Sensus Ekonomi

4.2 Results

High similarity examples:

Q1	Q2	Similarity
“Apa penghasilan utama rumah tangga Anda?”	“Berapa pendapatan utama keluarga Anda?”	0.89
“Apakah Anda bekerja minggu lalu?”	“Apa status pekerjaan Anda minggu lalu?”	0.83

Graph clustering yielded **12 clusters**, showing strong redundancy across directorates.

5. Discussion

The system effectively revealed:

- cross-survey duplication
- overlapping conceptual indicators
- potential survey instrument consolidation
- possibilities for metadata harmonization

Implications:

- reducing respondent burden
- improving statistical coherence

- enabling integrated indicator frameworks
- supporting modernization of BPS metadata management

6. Future Work

- multilingual IndoBERT + MiniLM hybrid embeddings
- hierarchical clustering
- topic modeling
- metadata knowledge graph
- federated deployment per directorate

7. Conclusion

This system demonstrates the feasibility of automatic detection of cross-survey redundancy using NLP-based semantic similarity.
It forms a foundation that can support BPS harmonization efforts and serves as a research-strengthening component for graduate study applications, including KAUST.