Data Preprocessing and Program Matching Workflow

Raw Weibo Posts
(88,253 posts)

→

Text Cleaning
Remove URLs, special chars

→

Chinese Segmentation
Jieba + Transit Dictionary

↓

TF-IDF Transform
Feature extraction

→

LDA Topic Modeling
K=15, α=0.05, β=0.005

↓

Program Descriptions
6 improvement programs

→

Neural Embeddings
MiniLM-L12-v2 (384-dim)

↓

Semantic Matching
Cosine similarity > 0.55

→

Matched Posts
17,618 posts (87.4% accuracy)

↓

Program-Specific Sentiment Analysis
Ready for ITSA evaluation

Key Processing Steps:

1. Text cleaning and Chinese word segmentation
2. TF-IDF feature extraction and LDA topic modeling
3. Neural embedding for semantic similarity matching
4. Threshold-based filtering for program relevance