CERIAS

The Center for Education and Research in Information Assurance and Security

Digital Guardian: Harnessing AI to Devalue False Information and Protect Public Discourse on Social Media

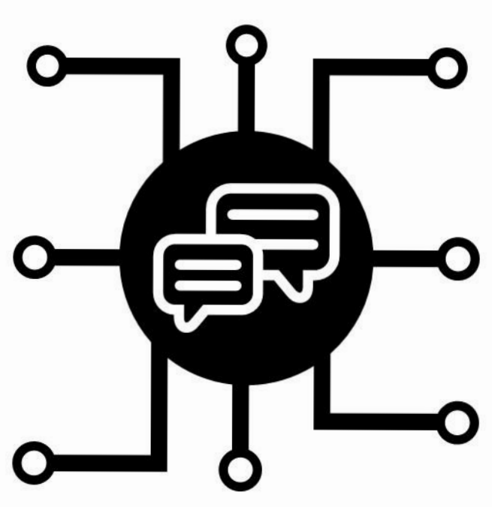
Authors: Nicholas Harrell, Information Security Ph.D. Student,
Affiliation: CERIAS, Purdue Military Research Initiative (PMRI), Purdue Homeland Security Institute (PHSI)

- Detecting potential false information that can become viral is an ongoing challenge.
- We demonstrate a method using human values as underlying components of information to detect potential virality.
- By leveraging Large Language Models (LLM) to generate a large synthetic dataset, we enhance viral signal detection in a Multi-Label Classification Model.

Real Live Data

Step 1. Collect Real Data from Specific Themes

Large Language Model



Step 2. Through Prompt engineering and fine-tuning, generate content, label pairs

Synthetic Data

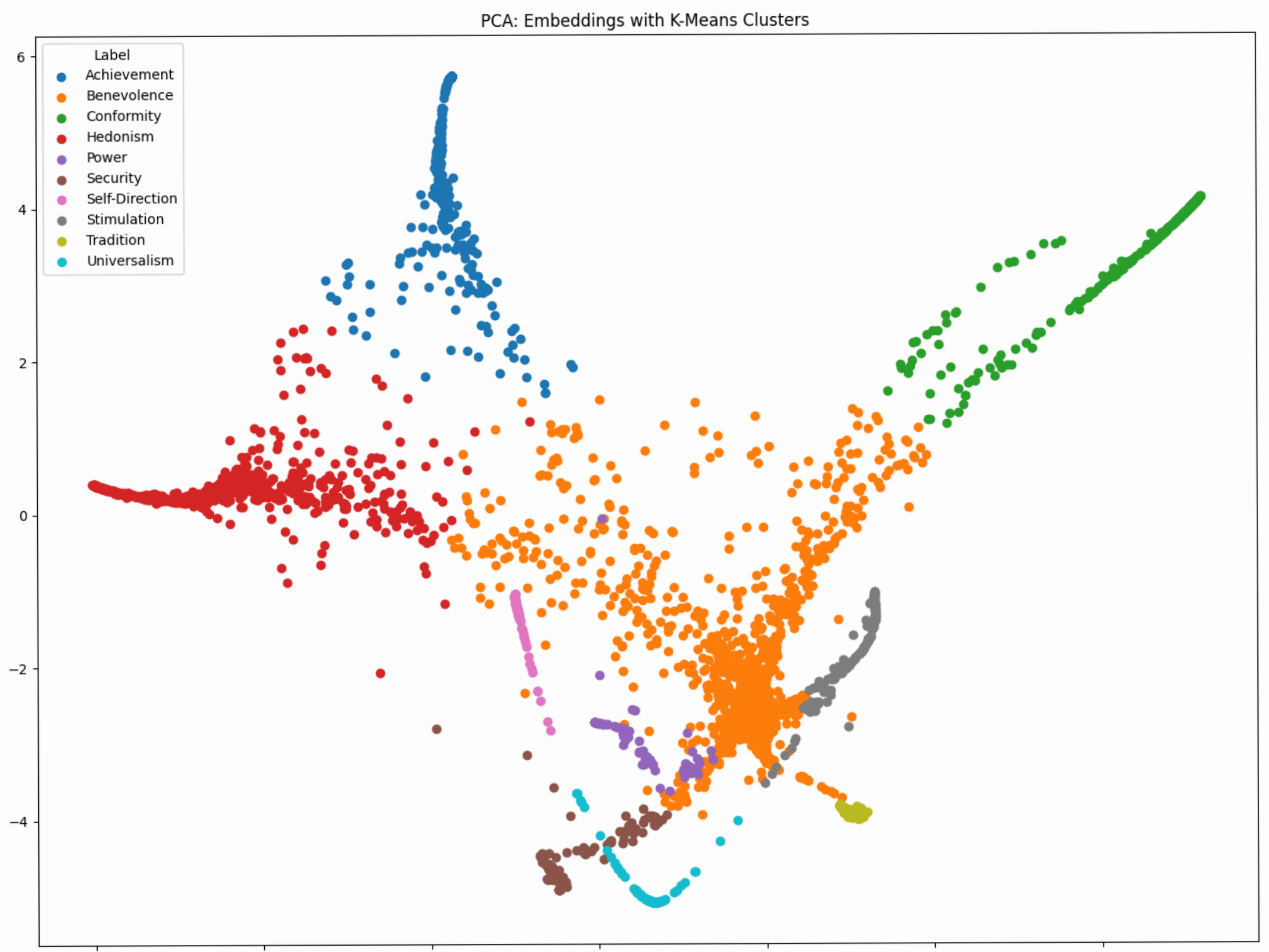
Step 3. Parse and Clean Dataset

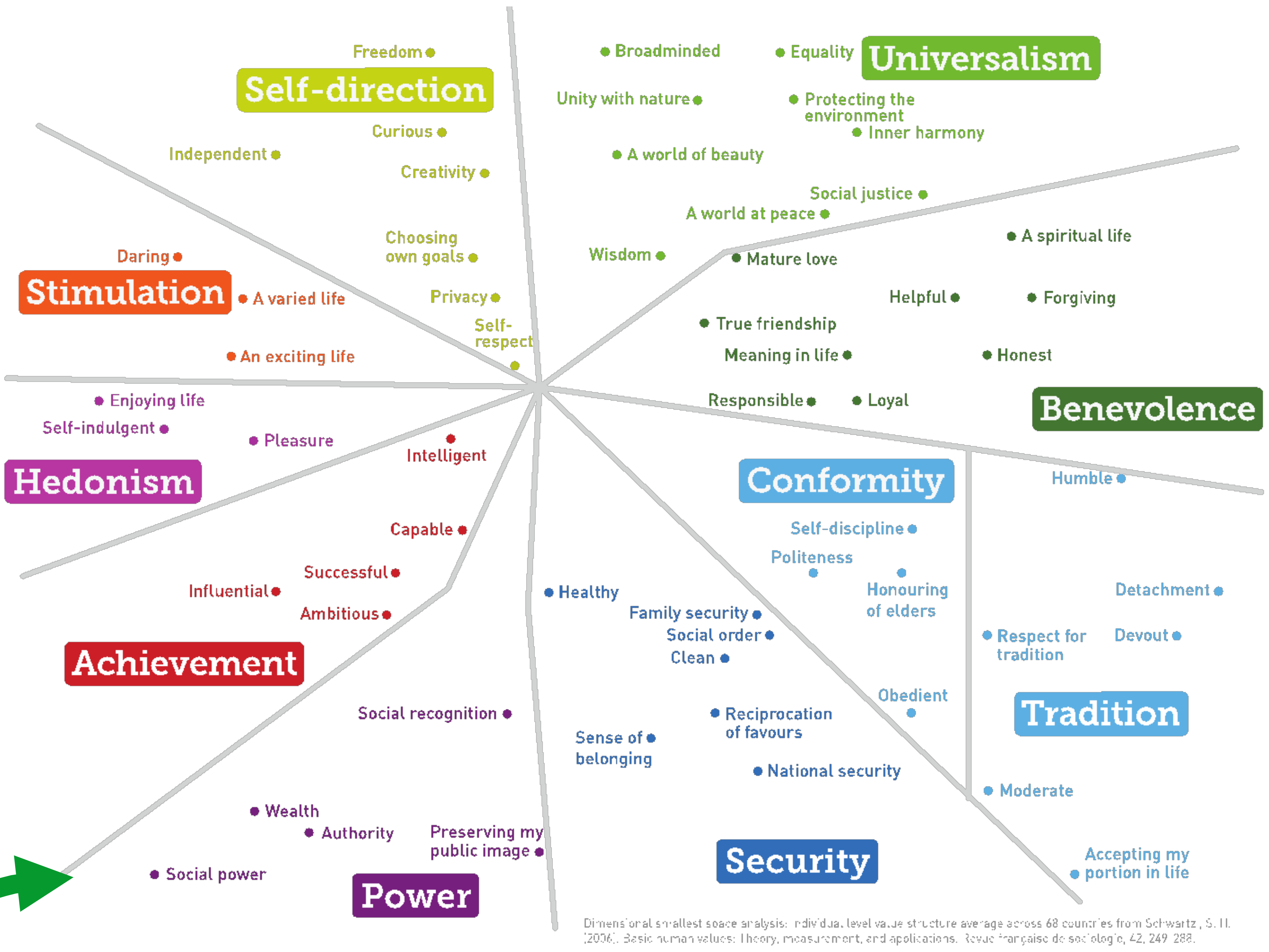
Multi-label Classifier

Step 3. Train Multi-label classifier with synthetic scenarios that map to values labels.

If weak correlations:
– return to Step 2

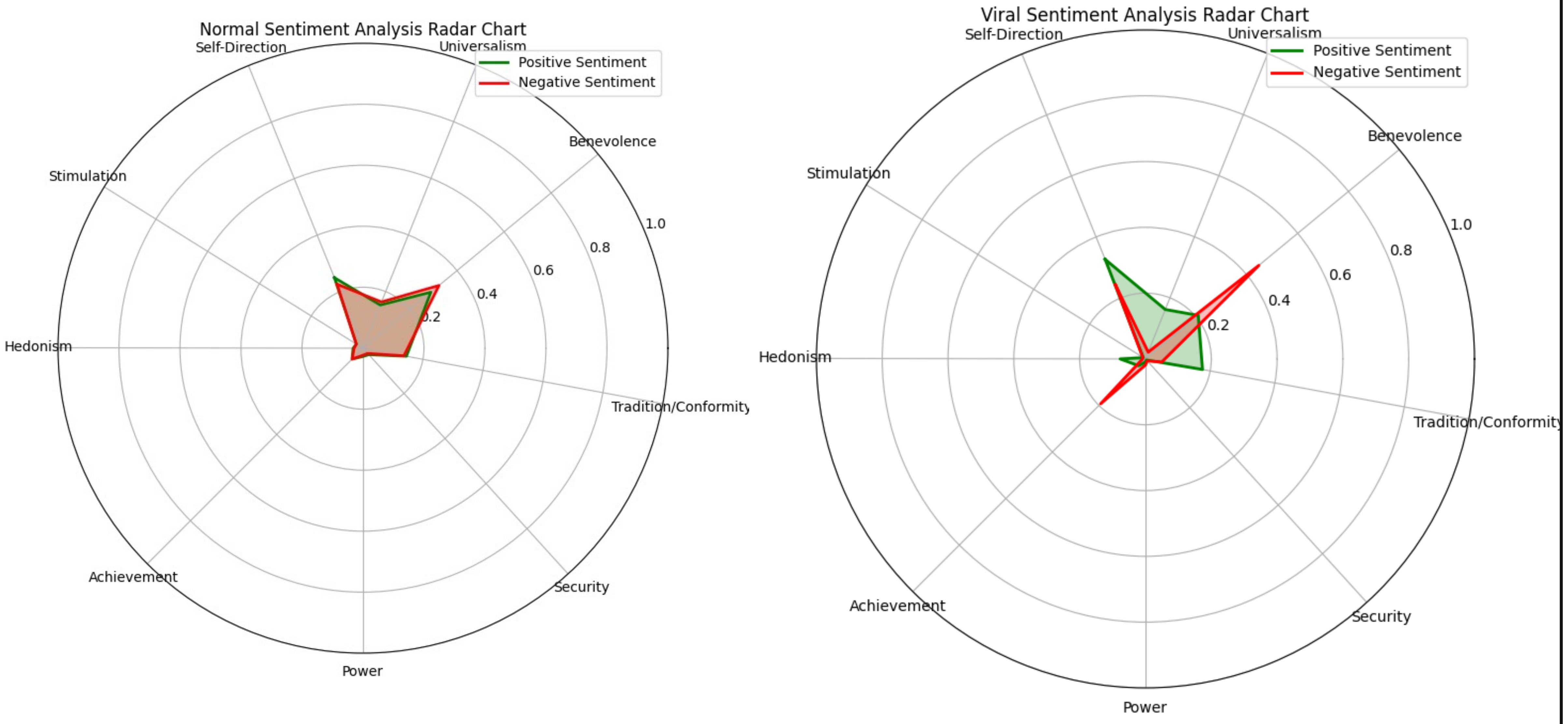
Step 4. Analyze loadings





Dimensional smallest space analysis: individual level value structure average across 68 countries from Schwartz, S. H. (2006). Basic human values: Theory, measurement, and applications. Revue française de sociologie, 42, 249 288.

Step 5: Measure effect between Normal Information and Viral Information



Viral Data shows a significant change in Benevolence and Achievement

Outcome: This method shows support for improving the detection of weak viral signals within social media text content.

PURDUE UNIVERSITY

CERIAS