# Information Extraction and Unsupervised Methods for Streamlining Evidence Synthesis in International Development Grey Literature

Kristen Edwards, Jack Gammack, Luis Kumanduri, Dylan Lewis
6.864 Final Project
Dec 7, 2021

# Background & Motivation

- ➔ Evidence based policy
- ➔ International development literature base
  - ◆ Vast and growing
  - ◆ Largely unstructured
- ➔ Evidence synthesis
  - ◆ Compiling all relevant information
  - ◆ Time and resource intensive


USAID
FROM THE AMERICAN PEOPLE

**Can we use Natural Language Processing to expedite manual evidence synthesis and uncover broad classes of documents in the international development field?**

# Information Extraction

# Country of Study (CoS) Extraction & Classification

## CoS Classification Accuracy for various Input and Model Combinations

| Text Input | Baseline: SSM | ESMS | ESMM | ESML | ESMT |
|---|---|---|---|---|---|
| **T** | 0.676 | 0.434 | 0.520 | 0.488 | 0.725 |
| **A** | 0.762 | 0.787 | 0.807 | 0.820 | 0.836 |
| **ID** | 0.139 | 0.0 | 0.0 | 0.0 | 0.0 |
| **OD** | 0.029 | 0.029 | 0.029 | 0.029 | 0.029 |
| **T + A** | 0.840 | 0.844 | 0.869 | 0.885 | **0.910** |
| **T + A + ID** | 0.832 | 0.828 | 0.861 | 0.873 | 0.893 |
| **T + A + OD** | 0.832 | 0.840 | 0.857 | 0.881 | 0.898 |
| **T + A + ID + OD** | 0.824 | 0.824 | 0.848 | 0.869 | 0.881 |

**Input Abbreviations:**

**T** = Title of Paper
**A** = Abstract of Paper
**ID** = Intervention Description of Paper
**OD** = Outcome Description of Paper
**X + Y** = Text field **X** concatenated with text field **Y**

**NER Model Abbreviations**

**Baseline:**
**SSM =** Simple Substring Matcher
**Pretrained SpaCy English NER Models:**
**ESMS** = CNN Model (one-hot word embeddings)
**ESMM** = CNN Model (word-vector table of 20k unique vectors (Dimension of 300) for ~500k words)
**ESML**= CNN Model (word-vector table of 685k unique vectors (Dimension of 300) for 685k words)
**ESMT**= RoBERTa-base Transformer Model

# Country of Study (CoS) Extraction & Classification

## Map of Predicted CoS


## Filtering by Predicted CoS

Country of Study (CoS) Predictions Filter Function

[61] Show code

CoS: Guatemala

[66] Show code

```
1. Title: A multicountry randomized controlled trial of co
   Abstract: Background: Reported benefits of maternal nut
2. Title: Zinc Absorption in Guatemalan Schoolchildren Fed
   Abstract: BACKGROUND: Poor bioavailability of zinc from
3. Title: Twice the Recommended Daily Allowance of Iron is
   Abstract: The effects of iron on zinc status, oxidative
```

# Document Clustering

# Group Clusters by Topic

- Use Google's Universal Sentence Encoder to capture semantic meaning of abstract, intervention, and outcome descriptions
- Cluster based on semantic meaning
- Show most common words within each topic using TF-IDF

**Intervention Description Example:** In a school milk intervention study conducted on Beijing girls aged 10 y at baseline, we showed that the subjects who received a 330-mL dietary milk supplement (milk fortified with calcium alone or with both calcium and vitamin D) on school days had greater increases in height (by 0.7-0.8%), sitting height (by 0.7-1.2%), total-body-size-adjusted bone mineral content (BMC; by 1.1-2.5%), and total-body bone mineral density (BMD; by 3.1-5.4%) after 2 y than did the unsupplemented control subjects.

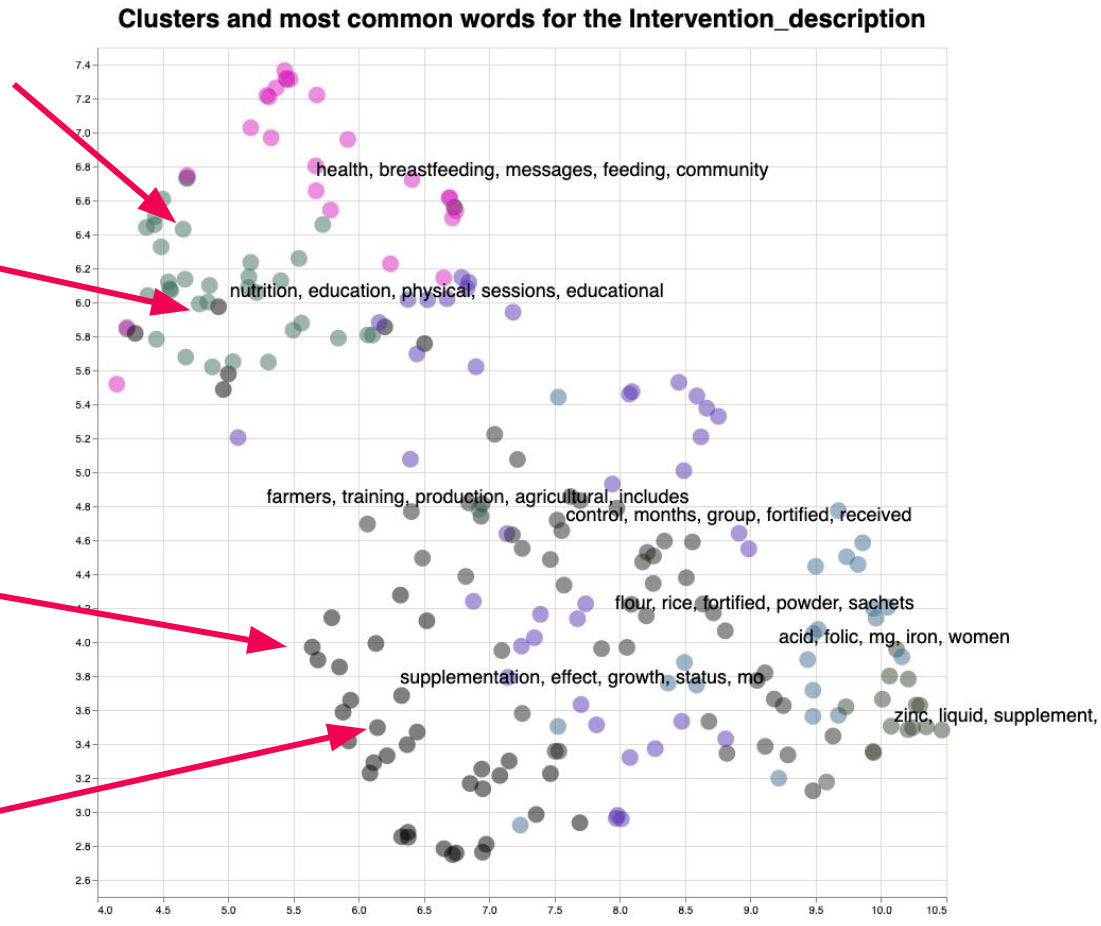
**Expert Labeled Category:** Direct provision of foods

"The intervention aimed to raise the profile of nutrition in the health facilities and to integrate nutrition services into existing child-oriented national programmes…"

"Nutrition education intervention was given to pregnant women between 1 and 4 months at baseline. )e education was given every 15 days for 5 consecutive months."

"We conducted a pilot randomized controlled trial to examine the effects of a 3-month supplementary feeding program delivered by community health workers on the nutritional status of mothers…"

"We examined the effects of preconception micronutrient supplementation on offspring growth and development with the use of data that were collected prospectively from offspring born to women…



# Analysis of Clustering

- Qualitatively, 7-8 clusters was most effective
- Clusters naturally correspond to categories of documents (e.g breastfeeding/micronutrients)
- Some bad clusters corresponding to common but irrelevant words, like “random” or “z-score”

# Conclusion

1. Transformer based NER model to predict country of study
2. Map-based UI for identifying geographically relevant documents
3. K-Means clustering to uncover distinct clusters of documents
4. Filtering capabilities using extracted information