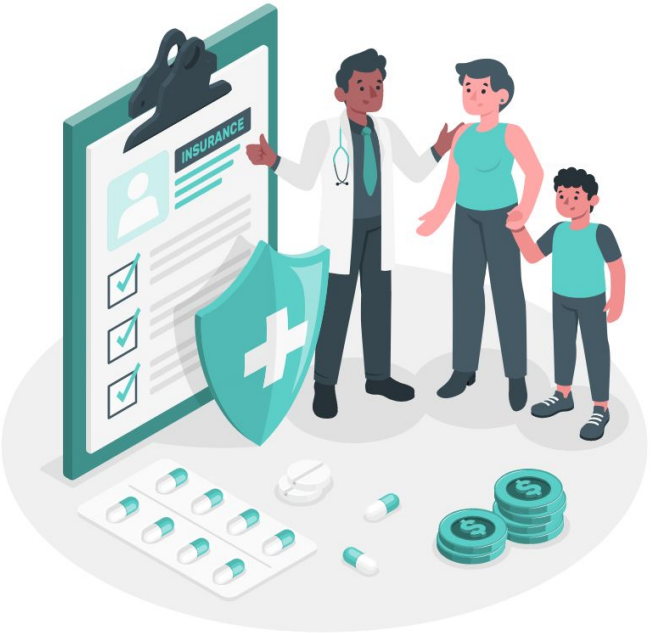
UN Sustainable Development Goal Text Classification

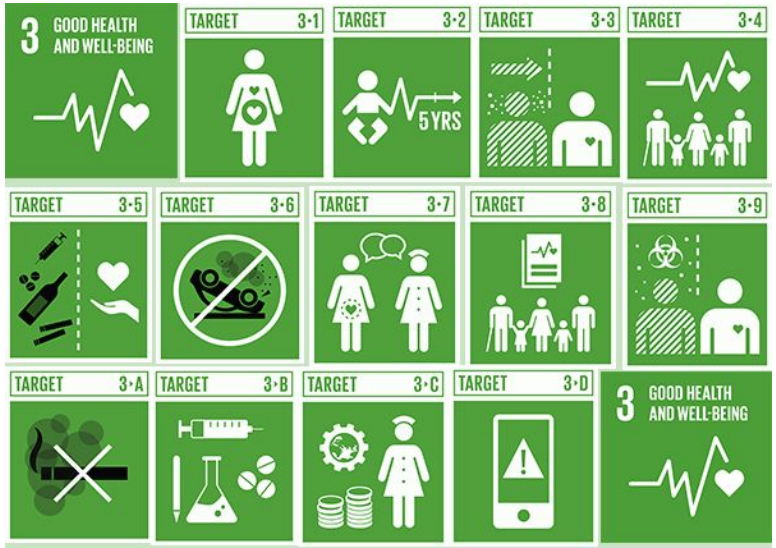


By. Suchanya (Mild) Trakarnsakdikul

What are the Sustainable Goals?

The United Nation created 17 measurable indication for development refer to as Sustainable Development Goals or refer to as SDG that they hope to achieve by 2030.

Our Focus



Goal 3: Good Health and Well Being

- 27 categories that provide quantitative measurement for Goal 3, example:
    - Maternal mortality ratio
    - Number of new HIV infections
    - Suicide mortality rate
    - Malaria incidence
- The categories are used as a judge to see whether the SDG has been achieved in 2030 or not

The Problem

Since there are many categories within the goal of Good Health and Well Being, the UN struggles to identify which contenders are related to which of the 27 categories.

Our Goal: Develope a text classifier to identify which contenders, using their text description, are most relevant to which categories

To predict which categories are relevant to the individual contenders

The Data

The Train Data

Includes **2995 web-scrapped** text with:

- **Unique ID:** IDs of text to be classified
- **Type:** Type of contenders
  - Grants
  - Tenders
  - Contract
  - News
  - Program
  - Organization
  - Open opportunity
  - Funding information
- **Text:** text to be classified
- **Labels:** Categories for each text (ie 3.1.1)

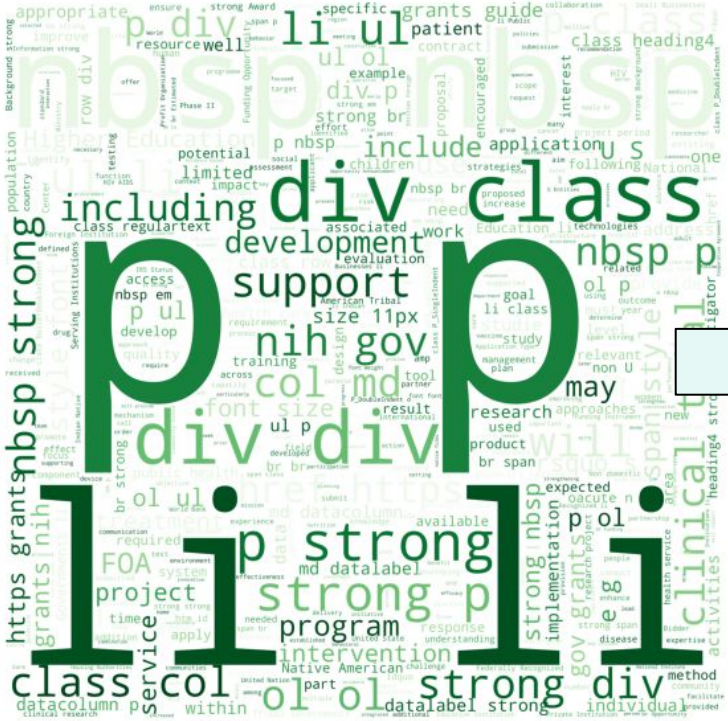
The Test Data

Includes **998 web-scrapped** text with:

- **Unique ID:** IDs of text to be classified
- **Type:** Type of contenders
  - Grants
  - Tenders
  - Contract
  - News
  - Program
  - Organization
  - Open opportunity
  - Funding information
- **Text:** text to be classified

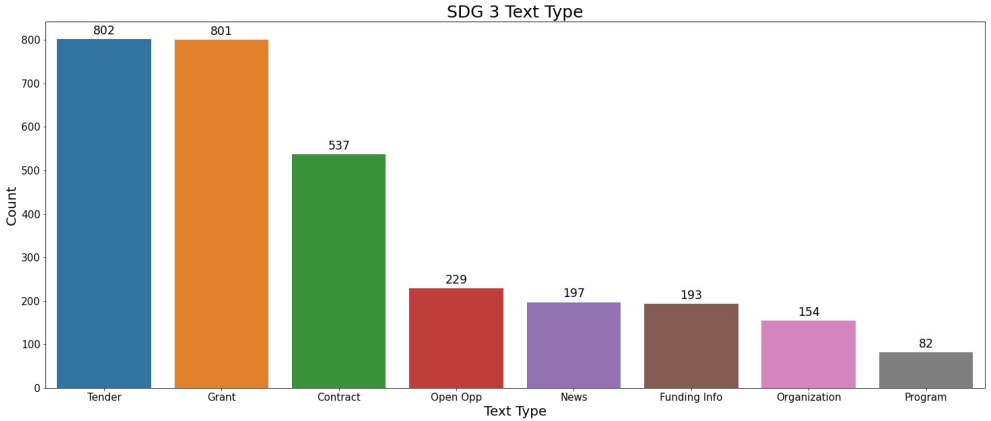
Data at a Glance

Original Data



Cleaned Data

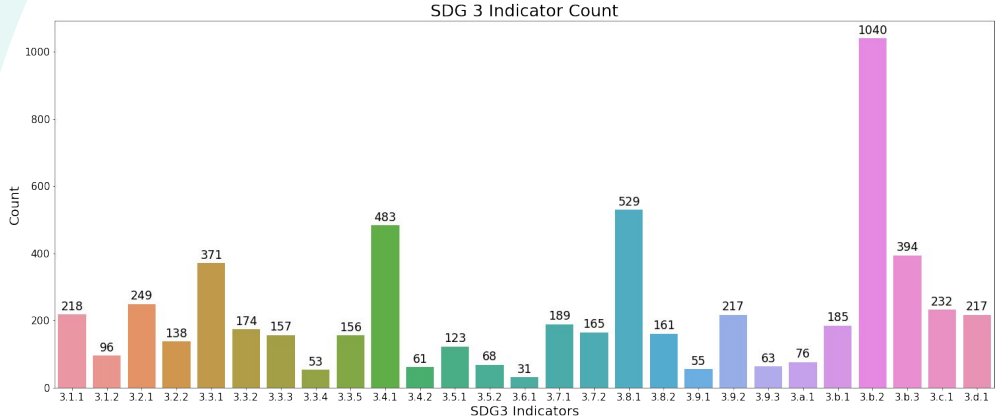
Understanding the Contender Type

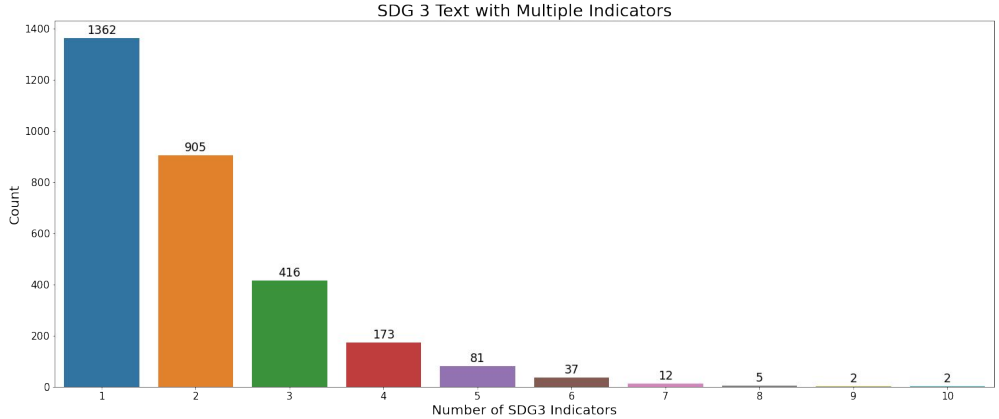


8 Types of Contender:

- **By Description**
  - Grant
  - Organization
  - Tender
  - Funding Information
  - Open Opportunity
  - Program
- **By Title or Headline**
  - Contract
  - News

Understanding the Categories



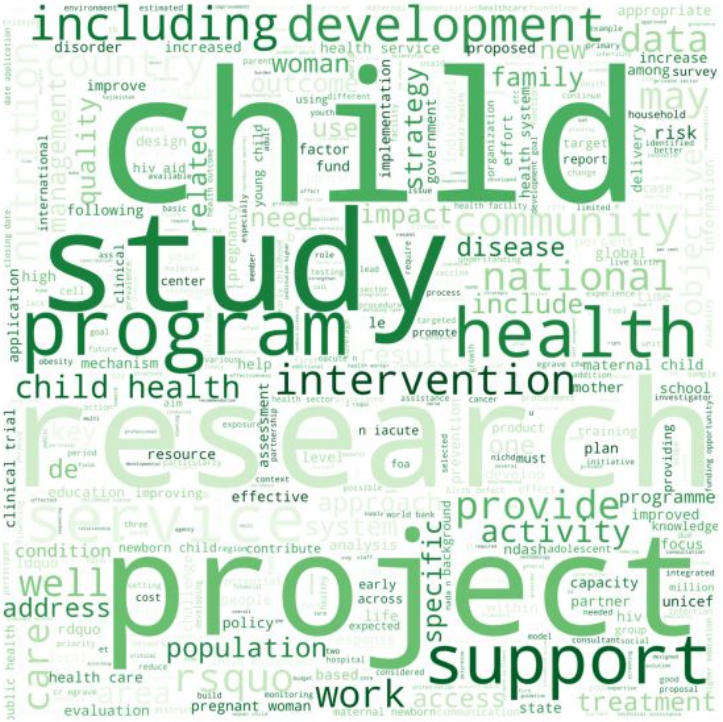


27 Categories:

- Most common categories:
  - 3.b.2 - International Health Regulation
- Least common categories:
  - 3.6.1 - Death rate due to road traffic injuries
- Most texts are classified by 1 and 2 categories

Text from Categories

Category 3.2.1:
Under-5 mortality rate

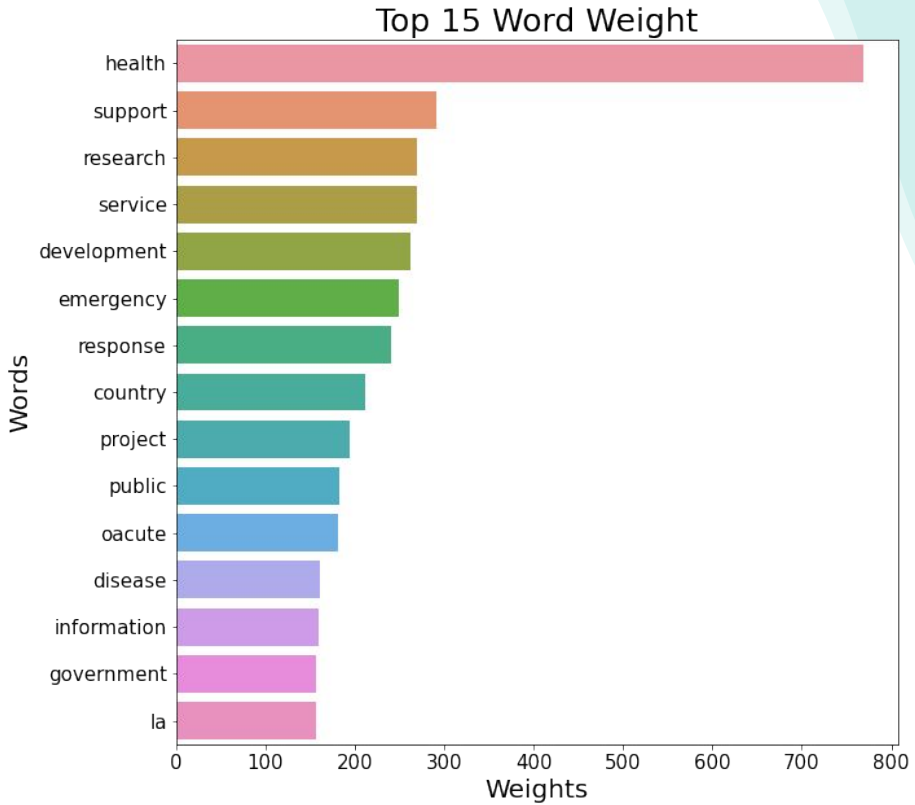


Category 3.3.2:
Tuberculosis incidence

Modeling and Prediction

KNN, MLP and Ridge Classifier

- TF-IDF Vectorizer to determined the value and weight of each words
- Fitted the best parameters for the estimators and vectorizers
- From the fitted parameter, an optimized model is created with high accuracy



Best Model

Ridge Classifier CV

- Use with MultiOutputClassifer function
- Use TF-IDF Vectorizer and ridge classifier default parameters
- 96% of the predicted categories were correctly predicted

Recommendation

The best text classifier to use:

Ridge Classifier

to help classify which tender is relevant
to which of the 27 categories

Additional Data

27 Unique Models

Optimized unique models for each individual categories

Stopwords Identification

Identify additional stop words for repeated vocabularies

Thank you

Appreciate your time

Appendix

The Categories

- The code and what it is

Data Visualization

- Word cloud for each categories
- Categories spread for type of data

27 Categories

There are 27 possible SDG 3 categories, in the dataset each indicator will be refer using code:

- 3.1.1 - Maternal mortality ratio
- 3.1.2 - Proportion of births attended by skilled health personnel
- 3.2.1 - Under-5 mortality rate
- 3.2.2 - Neonatal mortality rate
- 3.3.1 - Number of new HIV infections per 1 000 uninfected population, by sex, age and key populations
- 3.3.2 - Tuberculosis incidence per 100 000 population
- 3.3.3 - Malaria incidence per 1 000 population
- 3.3.4 - Hepatitis B incidence per 100 000 population
- 3.3.5 - Number of people requiring interventions against neglected tropical diseases

27 Categories

- 3.4.1 - Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease
- 3.4.2 - Suicide mortality rate
- 3.5.1 - Coverage of treatment interventions (pharmacological, psychosocial and rehabilitation and aftercare services) for substance use disorders
- 3.5.2 - Harmful use of alcohol, defined according to the national context as alcohol per capita consumption (aged 15 years and older) within a calendar year in litres of pure alcohol
- 3.6.1 - Death rate due to road traffic injuries
- 3.7.1 - Proportion of women of reproductive age (aged 15–49 years) who have their need for family planning satisfied with modern methods
- 3.7.2 - Adolescent birth rate (aged 10–14 years; aged 15–19 years) per 1 000 women in that age group

27 Categories

- 3.8.1 - Coverage of essential health services (defined as the average coverage of essential services based on tracer interventions that include reproductive, maternal, newborn and child health, infectious diseases, non-communicable diseases and service capacity and access, among the general and the most disadvantaged population)
- 3.8.2 - Proportion of population with large household expenditures on health as a share of total household expenditure or income
- 3.9.1 - Mortality rate attributed to household and ambient air pollution
- 3.9.2 - Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (exposure to unsafe Water, Sanitation and Hygiene for All (WASH) services)
- 3.9.3 - Mortality rate attributed to unintentional poisoning

27 Categories

- 3.b.2 - Total net official development assistance to medical research and basic health sector
- 3.b.3 - Proportion of health facilities that have a core set of relevant essential medicines available and affordable on a sustainable basis
- 3.c.1 - Health worker density and distribution
- 3.d.1 - International Health Regulations (IHR) capacity and health emergency preparedness
- 3.a.1 - Age-standardized prevalence of current tobacco use among persons aged 15 years and older
- 3.b.1 - Proportion of the target population covered by all vaccines included in their national programme

Word Cloud for 27 Categories

Categories Distribution by Text Type