

UN Sustainable Development Goal Text Classification

By. Suchanya (Mild) Trakarnsakdikul



TABLE OF CONTENTS



01

The United Nation

Getting to know who they are and their goals



02

The Data

What does the data looks like and what we need



03

Data Analysis

What can we gain from a closer look at the data



04

Modeling and Prediction

Exploring different model and the results from it



05

Recommendation

Exploring the best option and the future of the models



06

Thanks

Any questions at the end?



Getting to Know the United Nation

The Sustainable Development Goals

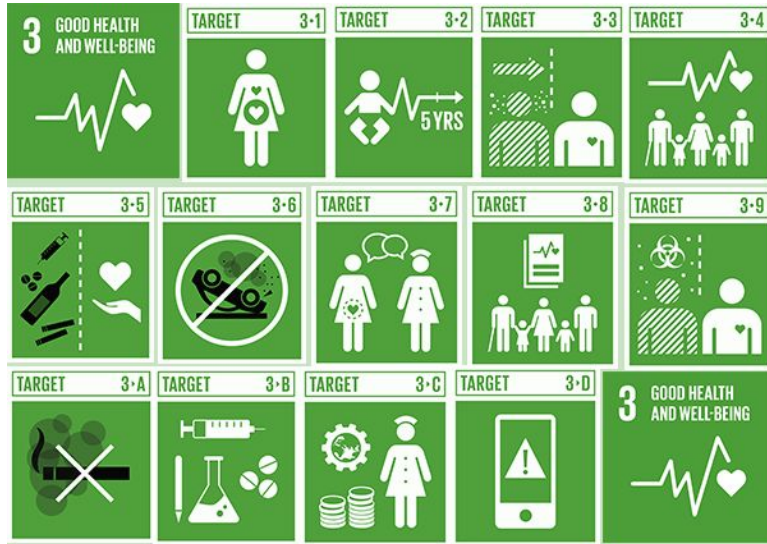


What are the Sustainable Goals?

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.



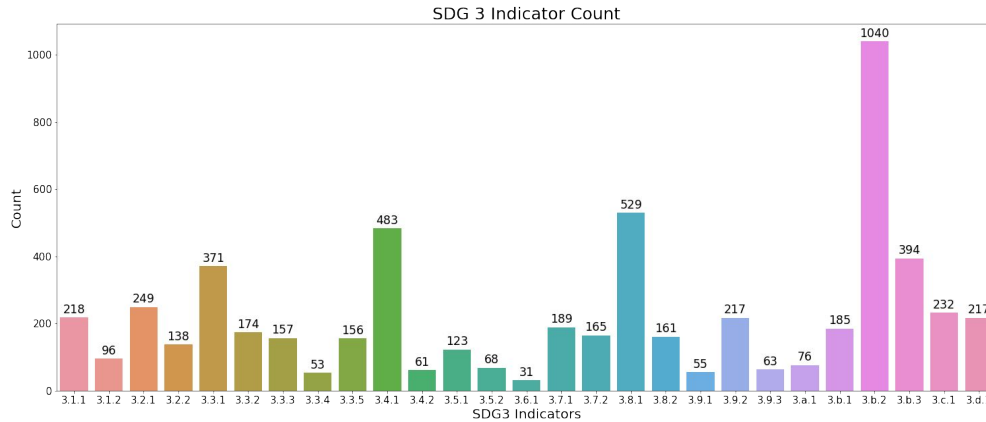
Our Focus



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

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

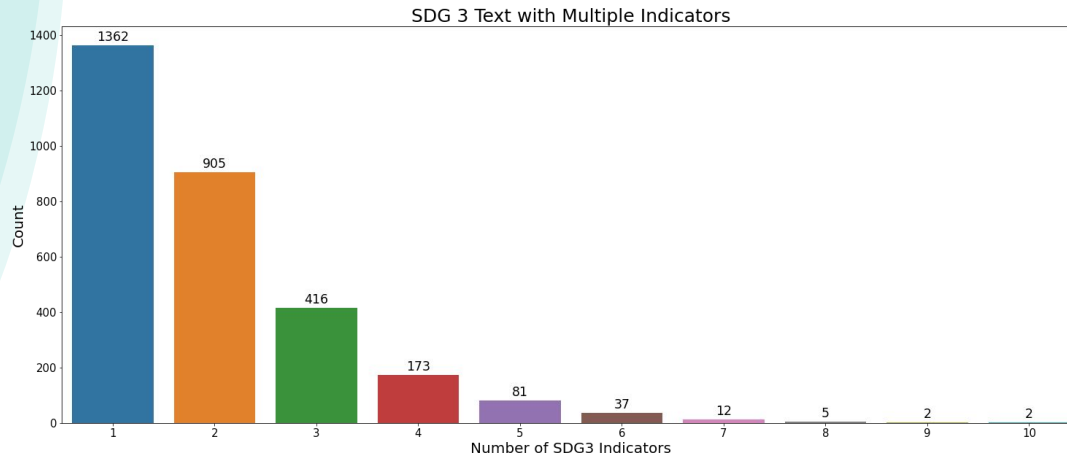
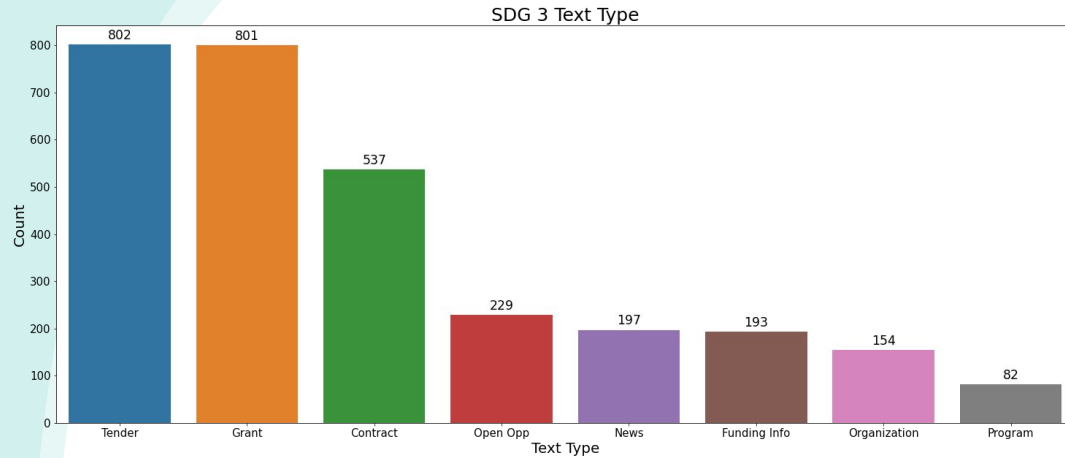
Since there are many categories, the UN needs to find contenders that can help aid each categories...

The Contenders

There are many contenders and international agencies that contributes to the sustainable development goals. These contenders can be in the form of organizations, funding, contracts, programs and other forms. Some notable examples are:

- World Health Organization (Organization)
- Cambodia Aid investment (Program)
- The Tropical Agriculture Association (Funding)





Our Contenders

8 Types of Contender:

- Grant
- Organization
- Tender
- Funding Information
- Open Opportunity
- Program
- Contract
- News

One contenders can be relevant to more than one categories:

- Most are related to 1 to 3 categories

The Problem

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

Our Goal: Develop 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

What are we exploring



The Data



The Train Data

Includes **2995 web-scraped** 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-scraped** 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

Text Example - Description

grant_text.Text[50]

'Evaluating Policies for Impacts on Multiple Forms of Violence <p>Background: </p> <p>Violence is a significant public health problem in the United States. In 2015, more than 62,000 people in the United States died because of violence and more than 2,165,000 were treated in emergency departments for a violence-related injury (CDC, 2016a). Exposure to violence in childhood and adolescence can increase risk for later violent experiences, such as intimate partner violence, sexual violence, and suicide, which can have a cumulative and compounding impact on health and well-being. This is alarming given the prevalence of violence among children and youth. In 2015, approximately 1,670 children died in the United States from child abuse and neglect, and approximately 683,000 children were victims of child abuse and neglect per Child Protective Services (U.S. Department of Health & Human Services, 2017). Youth violence is also prevalent among persons aged 10 to 24 years; each day approximately 13 young people are victims of homicide and more than 1,300 are treated in emergency departments for nonfatal physical assault-related injuries (CDC, 2016a). Additionally, 1 in 5 high school students reported being bullied at school or getting in a physical fight in the past year (CDC, 2016b). Among high school students who reported dating during the 12 months before the survey, approximately 10% experienced physical dating violence and approximately 11% experienced sexual dating violence one or more times during the 12 months before the survey (CDC, 2016b). These forms of violence are common across the lifespan, with approximately 37% of U.S. women and 31% of U.S. men experiencing sexual violence, physical violence, and/or stalking by an intimate partner in their lifetime (Smith et al., 2017). In the U.S., approximately 1 in 3 women and nearly 1 in 6 men experience some form of contact sexual violence in their lifetime (Smith et al., 2017). In 2015, 44,193 individuals died by suicide, and between 1999 and 2015 suicide rates increased 27% (CDC, 2016a). According to self-report survey data, 1.4 million adults attempted suicide, 2.7 million made plans for suicide, and 9.8 million adults seriously considered suicide in 2015 (Center for Behavioral Health Statistics and Quality, 2016).</p>'

org_text.Text[110]

"TAA Agribusiness Group: <p>The Tropical Agriculture Association (TAA) is a professional association of individuals and corporate bodies concerned with the role of agriculture for development throughout the world. TAA brings together individuals and organizations from both developed and less developed countries to enable them to contribute to international policies and actions aimed at reducing poverty and improving livelihoods.</p> <p> </p> <p>The Association was formed in the late 1970's and stemmed largely from the alumni of the Imperial College of Tropical Agriculture, Trinidad. Over the years its membership has broadened to include all those interested in the various aspects of agricultural development worldwide. It has produced a quarterly Newsletter since 1981 that is now called 'Agriculture for Development'.</p> <p> </p> <p>Mission</p> <p>To advance education, research and practice in agriculture* for development</p> <p> </p> <p>Association's Primary Objective</p> Contribute to international policies aimed at reducing poverty and improving livelihoods in rural areas in the tropics, sub-tropics and countries with less developed economies in temperate areas. Encourage efficient and sustainable use of local resources and technologies, to arrest and reverse the degradation of the natural resources base on which agriculture depends, and to raise productivity of both agriculture and related enterprises to increase family incomes and commercial investment in the rural sector. <p> </p> <p>Particular emphasis is given to rural areas in the tropics and subtropics and to countries with less-developed economies in temperate areas. TAA recognizes the interrelated roles of farmers and other stakeholders living in rural areas, scientists (agriculturalists, economists, sociologists, etc.), government and the private sector in achieving a convergent approach to rural development. This includes recognition of the importance of the role of women, the effect of AIDS and other social and cultural issues on the rural economy and livelihoods.</p> <p> </p> <p>Services</p> Seminars/meetings on key issues in agriculture and development. Visits to organisations, companies and other interesting venues. Social events. "

Text from:

- Grants
- Organizations
- Tenders
- Funding Information
- Open Opportunity
- Programs

These text consist of full description with an average of 700 to 1200 words

Text Example - Titles

```
news_text[['word_count', 'unique_
```

	word_count	unique_word
count	197.000000	197.000000
mean	10.598985	10.487310
std	2.481893	2.391757
min	5.000000	5.000000
25%	9.000000	9.000000
50%	10.000000	10.000000
75%	12.000000	12.000000
max	18.000000	18.000000

```
contract_text[['word_count', 'uni
```

	word_count	unique_word
count	537.000000	537.000000
mean	9.538175	9.067039
std	5.306548	4.599154
min	2.000000	2.000000
25%	5.000000	5.000000
50%	9.000000	8.000000
75%	13.000000	12.000000
max	36.000000	27.000000

Text from:

- News
- Contract

These text consist of titles and headlines with an average of 10 words

```
contract_text.Text[2976]
```

'Equitable Access to Diabetes Care'

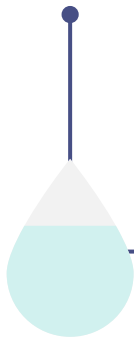
```
news_text.Text[21]
```

'Applying a workplace model to family planning outreach in the Philippines:'

Text Preprocessing

To identify which categories the contenders are relevant in, we will be exploring their descriptions of who they are and what they do. First we have to clean the texts:

**Removing
HTML Marking**



Anything behind
https://

Anything with

**Removing
Empty Spaces**



**Removing
Non-letters**



Anything with
numbers, symbols

Any repeating and
stopwords

**Removing
Stopwords**



Original Data

Cleaned Data

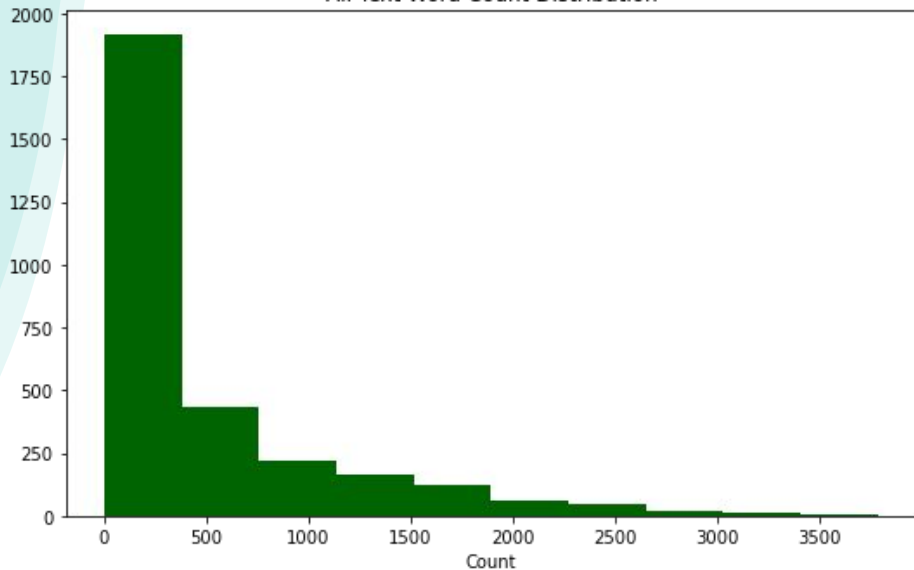


Features Engineering

Word Count

Count number of words in the contender description text

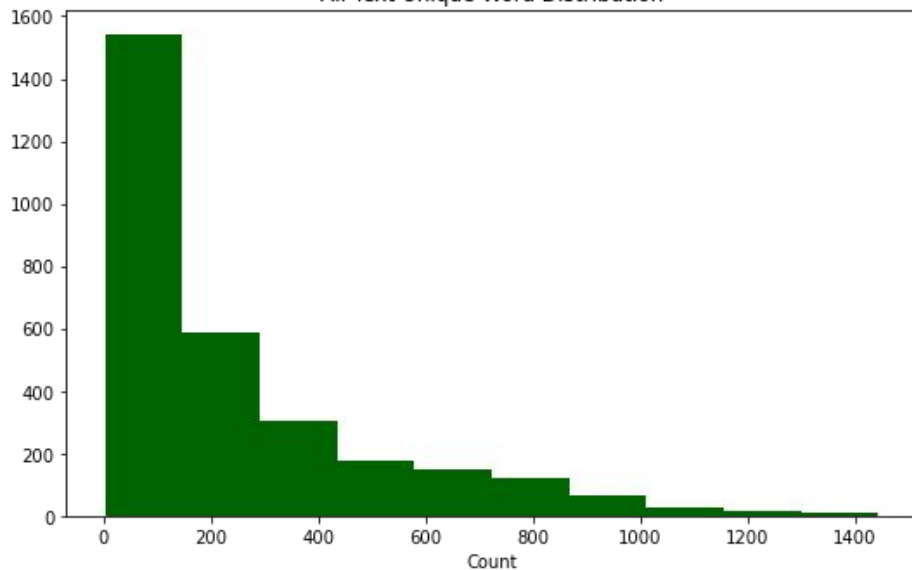
All Text Word Count Distribution



Unique Word Count

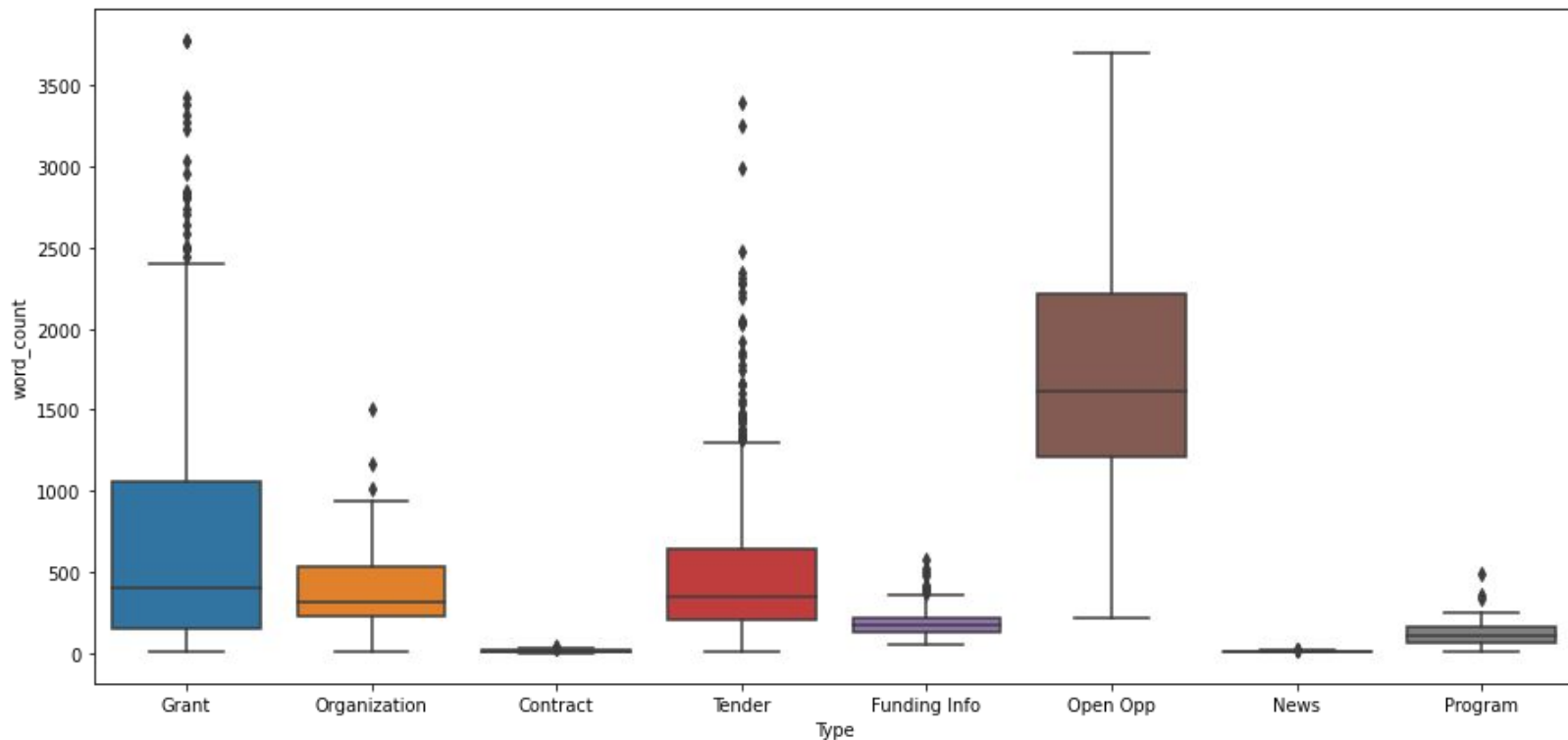
Count number of unique words in the contender description text

All Text Unique Word Distribution



The Text Spread by Text Type

Most of the text are longer as they are full descriptions of who the contenders are and what they do



Taking a Closer Look

Exploring Text Type and Categories

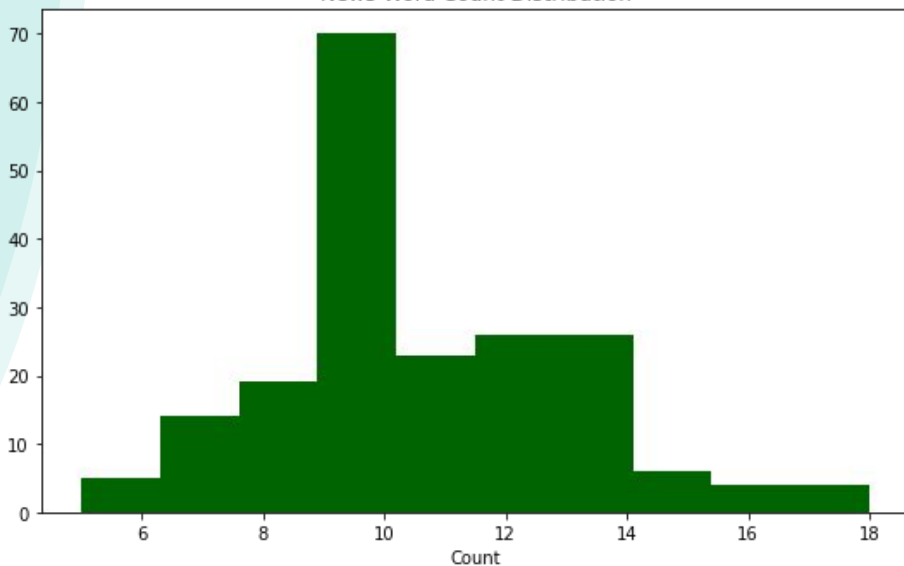


Looking Closer at the Text Type

Different text type have different spread in word count:

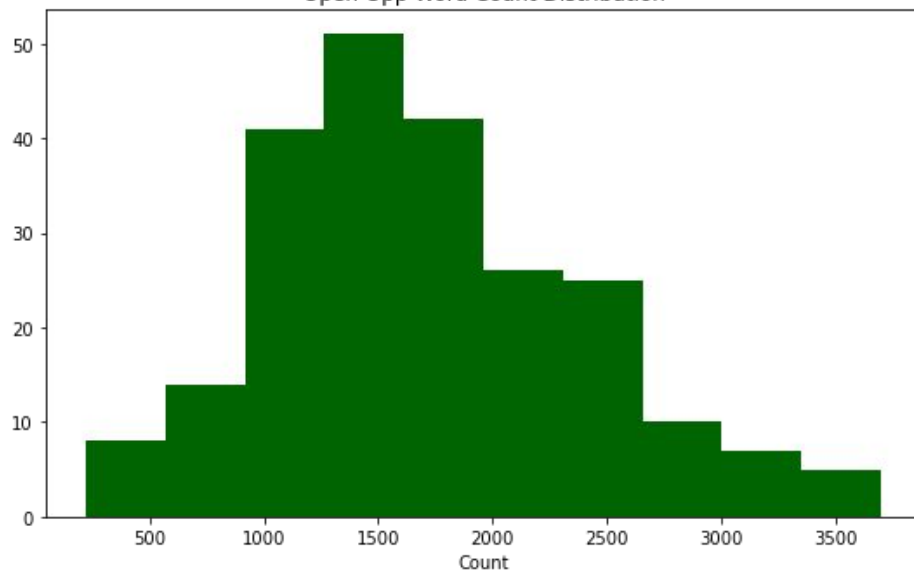
News Contender

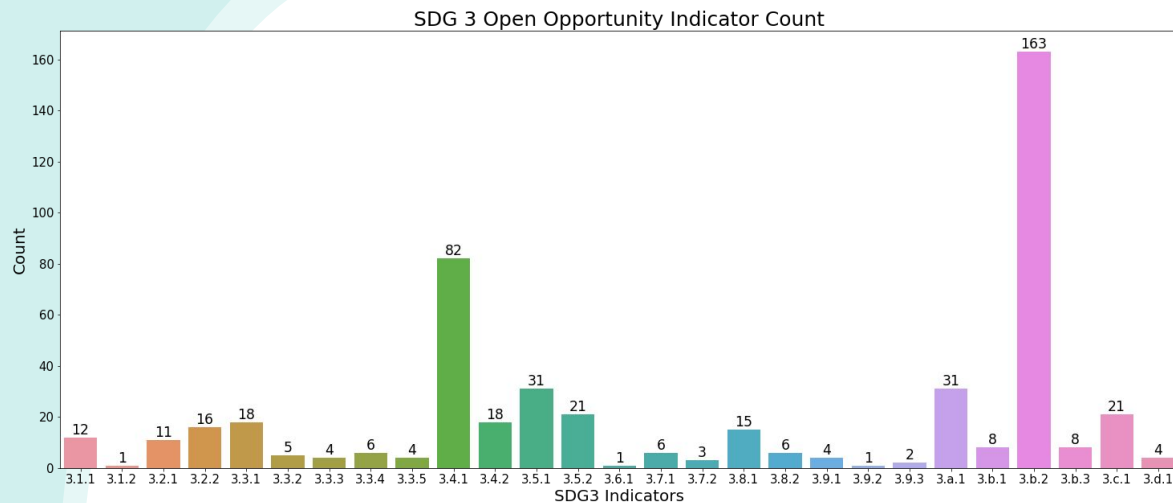
News Word Count Distribution



Open Opportunity Contender

Open Opp Word Count Distribution





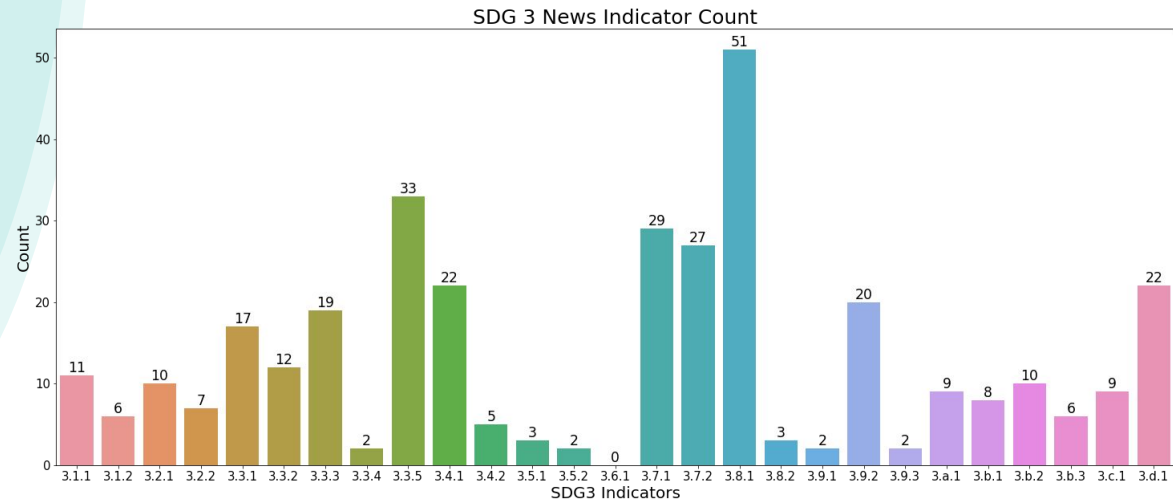
Category Count by Text Type

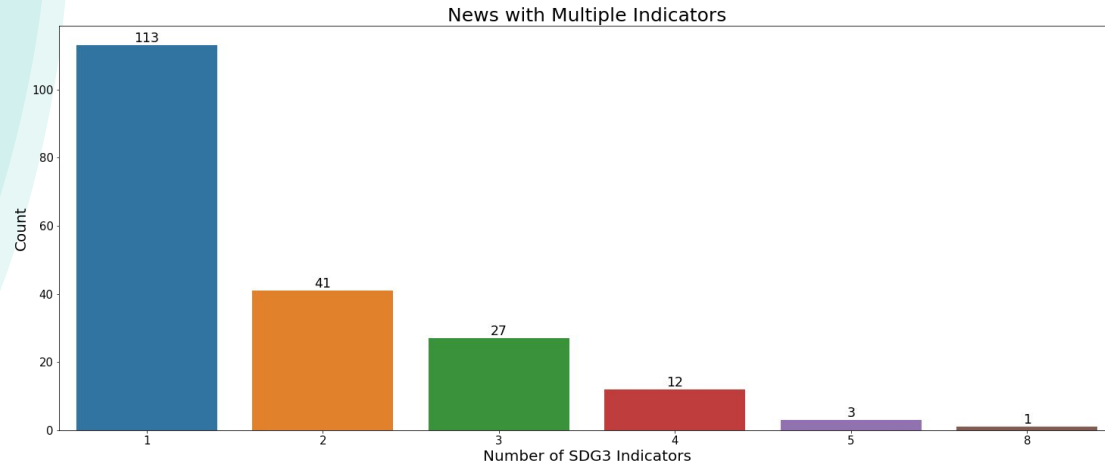
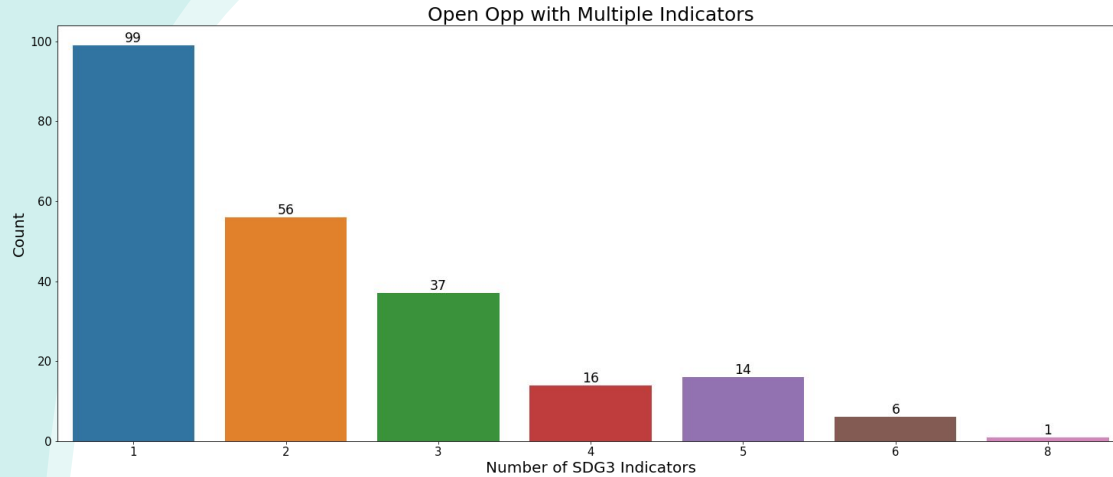
Open Opportunity

- Top Category: 3.b.1
- Low Category: 3.1.2, 3.6.1

News

- Top Category: 3.8.1
- Low Category: 3.6.1, 3.3.4





Number of Category by Text Type

Funding Information

- Most relevant between 1 to 3 categories
- Range from 1 to 8 relevant categories

News

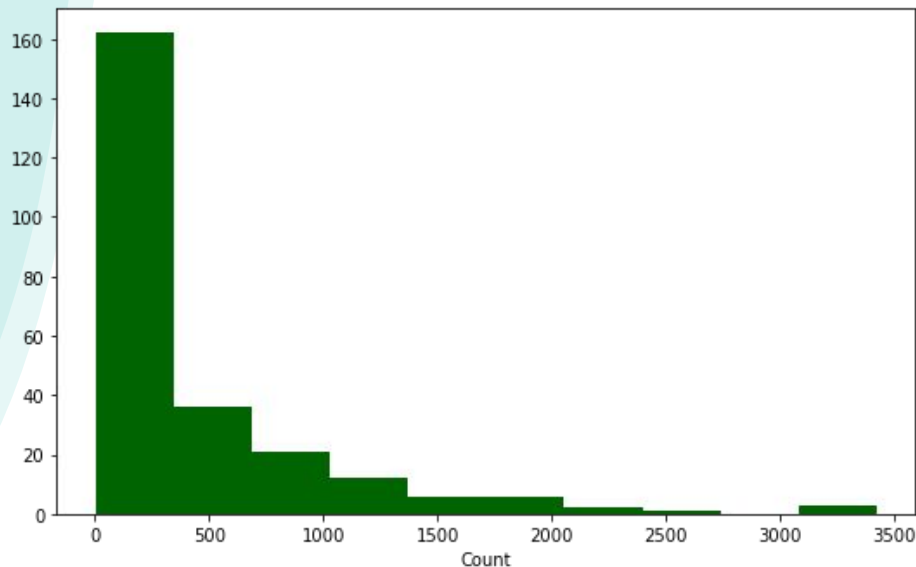
- Most relevant to 1 categories
- Range from 1 to 5

Looking Closer at the Categories

Different Categories have different spread in word count:

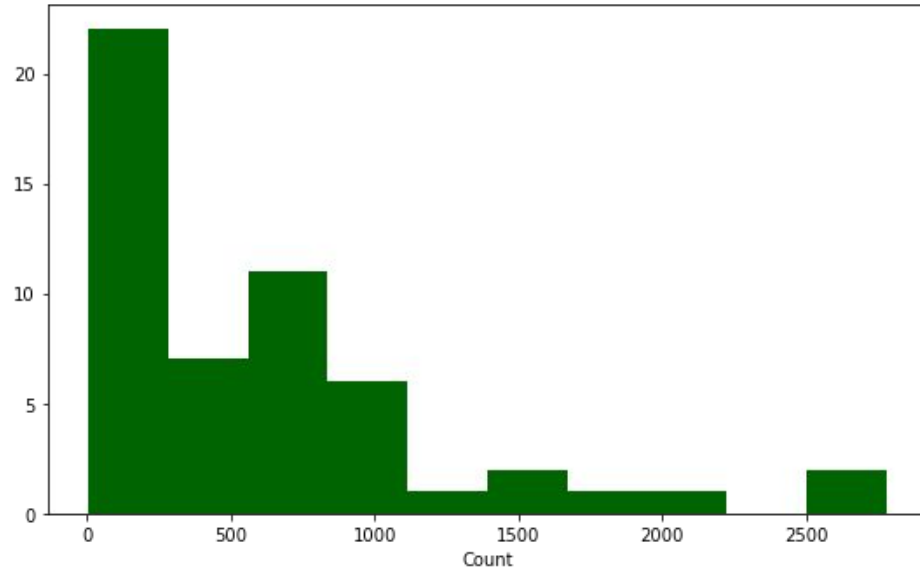
Category 3.2.1: Under-5 mortality rate

3.2.1 Word Count Distribution



Category 3.3.2: Tuberculosis incidence

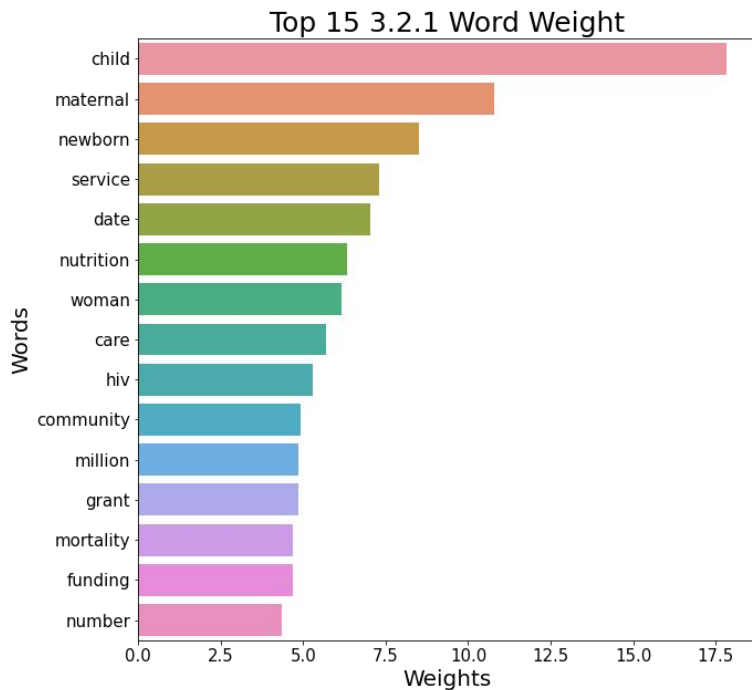
3.3.2 Word Count Distribution



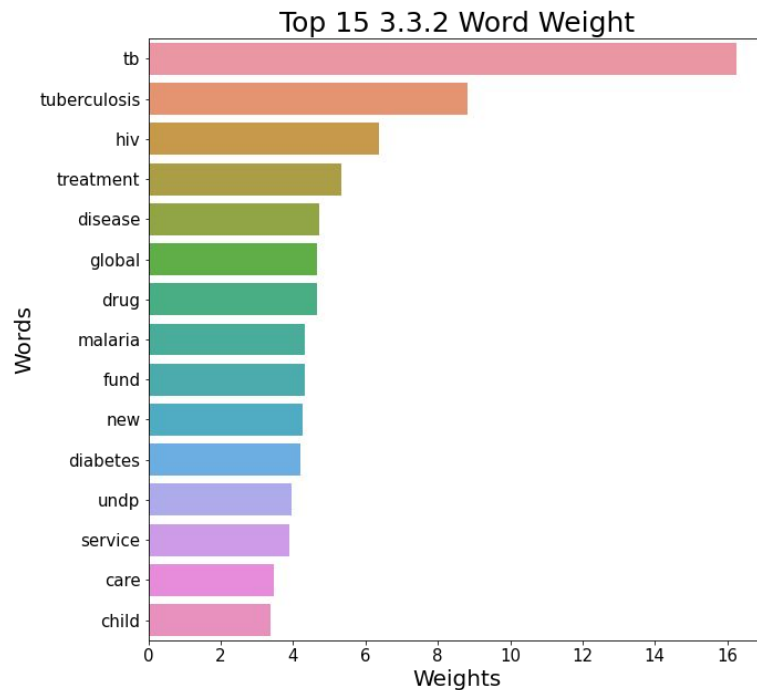
Text from Categories

Each categories contender text description emphasis on different things according to the categories example:

Category 3.2.1: Under-5 mortality rate



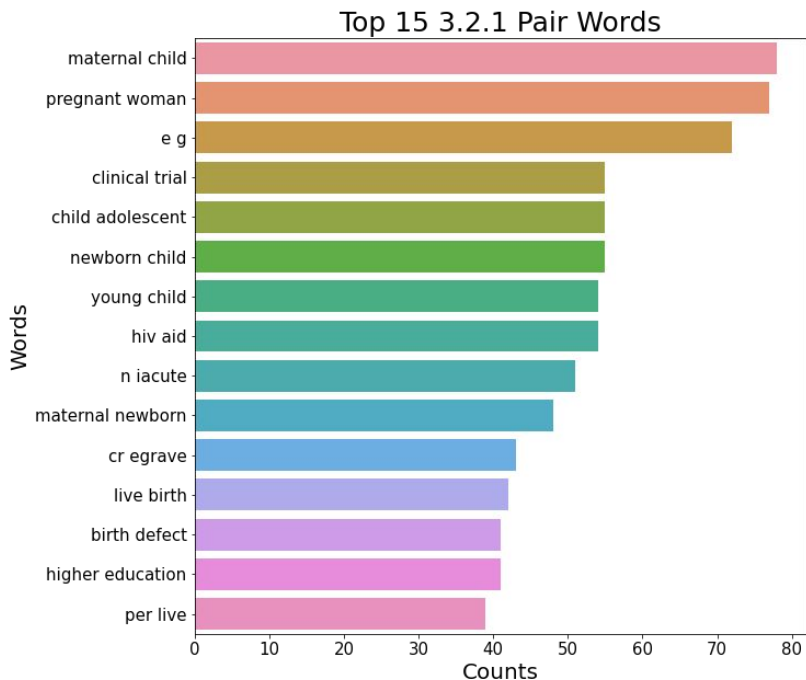
Category 3.3.2: Tuberculosis incidence



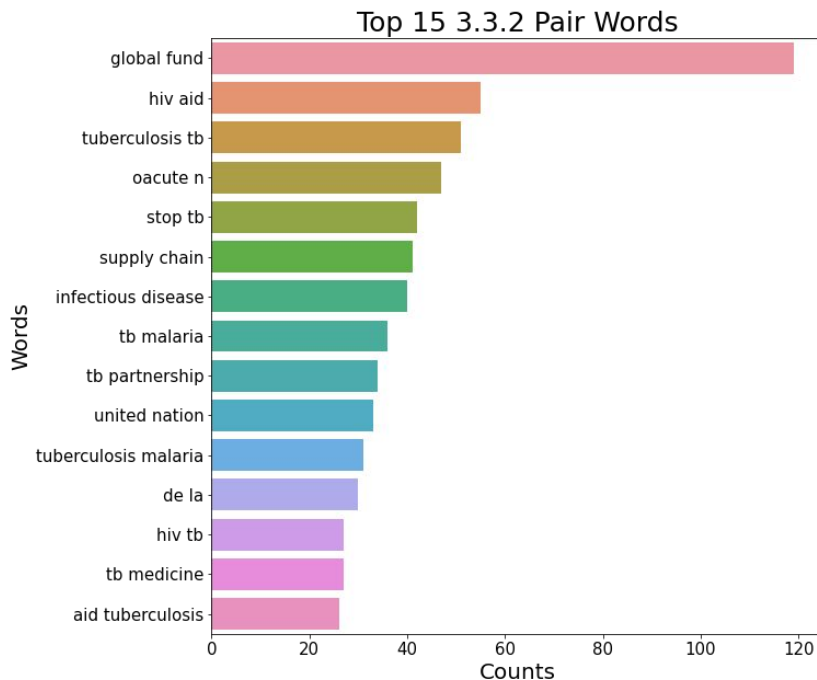
Pair Words from Categories

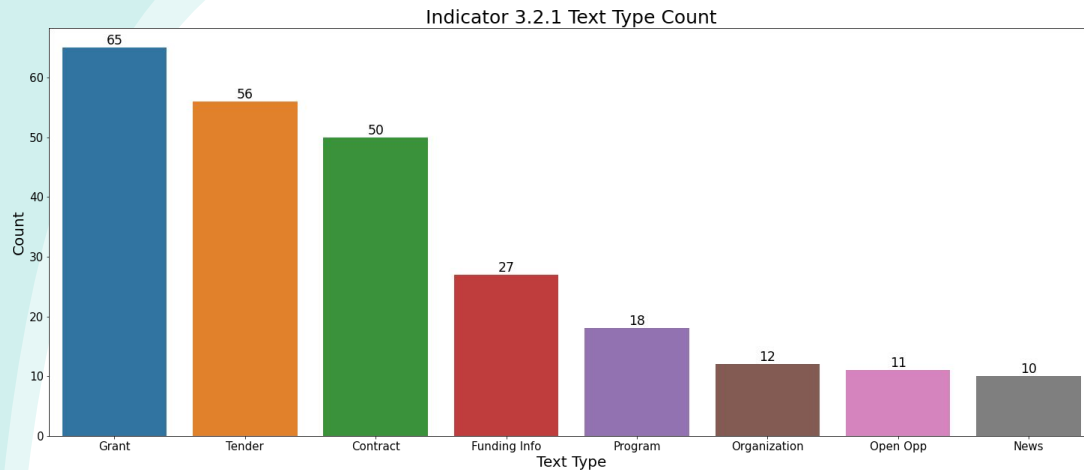
Each categories contender text description emphasis on different things according to the categories example:

Category 3.2.1: Under-5 mortality rate



Category 3.3.2: Tuberculosis incidence

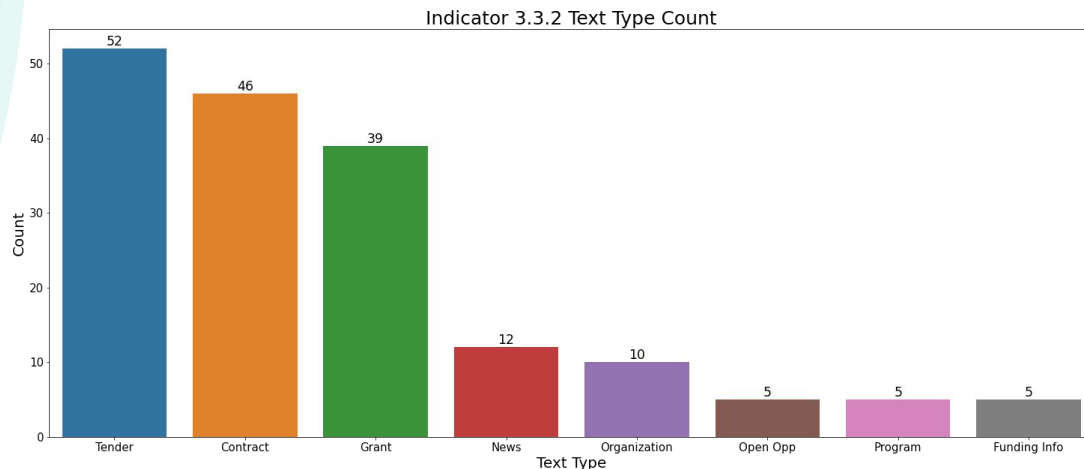




Text Type from Categories

Category 3.2.1: Under-5 mortality rate

- Grant (65)
- Tender (56)
- Contract (50)



Category 3.3.2: Tuberculosis incidence

- Tender (52)
- Contract (46)
- Grant (39)

Correlation?

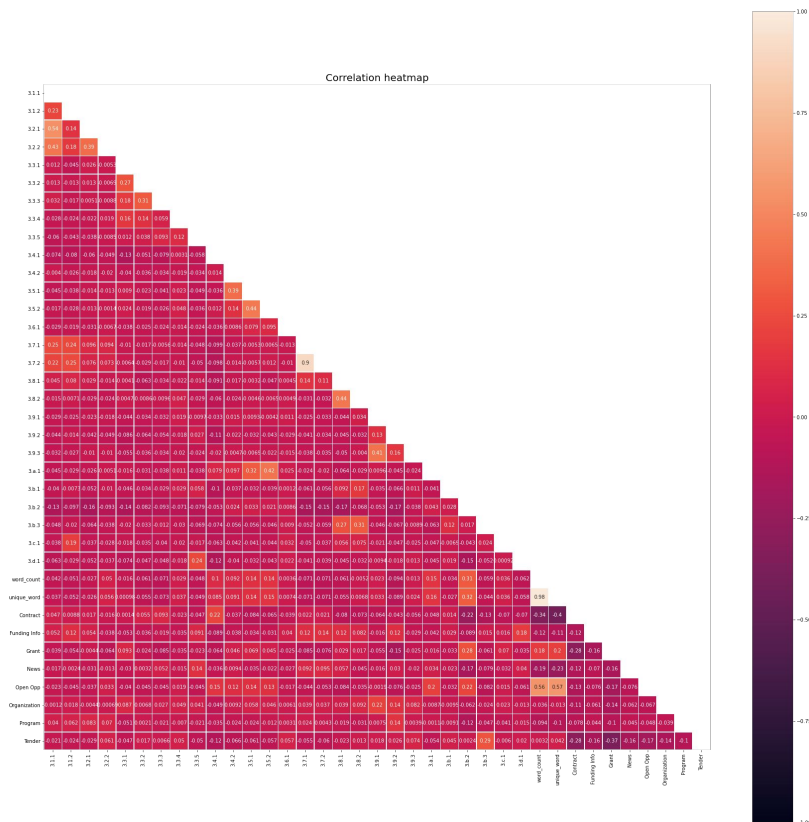
Overall, there is **no correlations** between...

- Type of contenders or word counts of the contenders description

VS

- Any of the categories

Shown by the colour heat map →



Modeling and Prediction

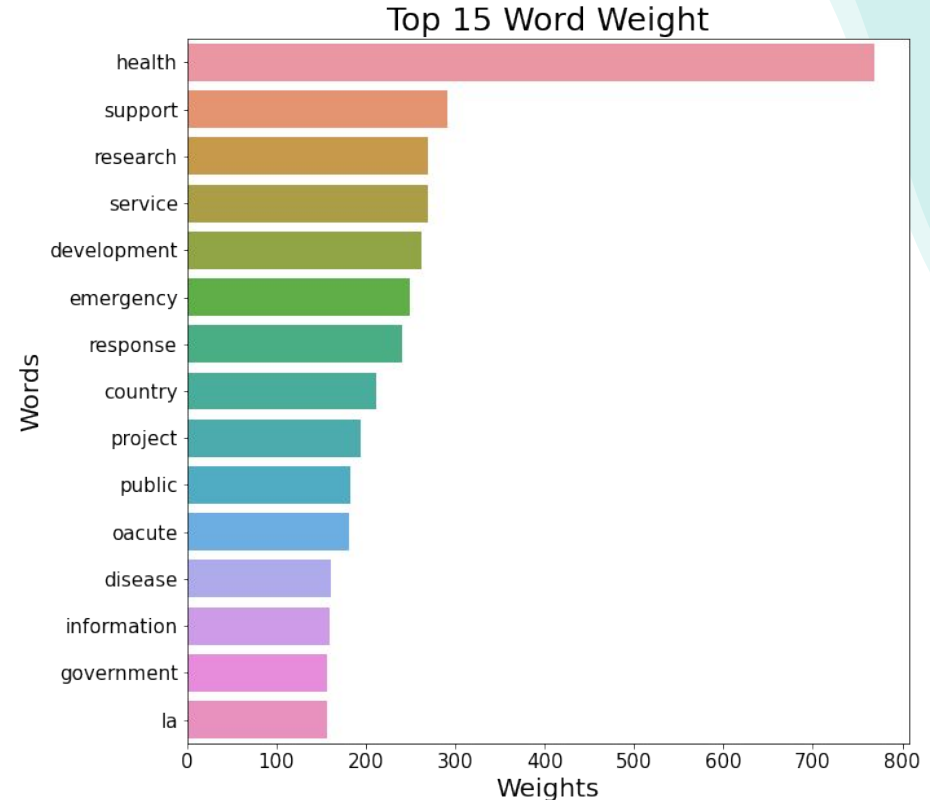
Putting our data to use



The Top Three Model

KNN, Random Forest and Ridge Classifier

- Using only the text data as the input
- TF-IDF Vectorizer to determined the value and weight of each words
- From the fitted parameter, an optimized model is created with high accuracy



From the Top Three

Evaluating the models by:

Hamming Loss Score which is the fraction of labels that are incorrectly predicted

Accuracy or the **Exact Match Ratio** which is the most struct metrics, indicating the percentage of samples that have all their labels classified correctly.

KNN Classifier

- Hamming Loss: 0.0553
- Accuracy: 0.2921

Random Forest

- Hamming Loss: 0.0555
- Accuracy: 0.2604

Ridge Classifier

- Hamming Loss: 0.0472
- Accuracy: 0.327



Product Model

Ridge Classifier CV

- Based on Ridge regression method, use with MultiOutputClassifier function
- Use TF-IDF Vectorizer and ridge classifier default parameters
- 96% of predicted categories were correctly predicted
 - Hamming Loss Score of 0.04
- Cross Validation Score
 - R2 Score of 0.62
 - F1 weighted Score of 0.52

Product Model Evaluation

	precision	recall	f1-score	support
'3.1.1'	1.00	0.43	0.60	30
'3.1.2'	0.00	0.00	0.00	18
'3.2.1'	0.86	0.40	0.55	45
'3.2.2'	0.67	0.11	0.18	19
'3.3.1'	0.96	0.63	0.76	78
'3.3.2'	0.96	0.49	0.65	47
'3.3.3'	0.96	0.53	0.69	45
'3.3.4'	1.00	0.24	0.38	17
'3.3.5'	0.72	0.38	0.50	34
'3.4.1'	0.93	0.56	0.70	93
'3.4.2'	0.75	0.27	0.40	11
'3.5.1'	0.78	0.44	0.56	16
'3.5.2'	0.83	0.33	0.48	15
'3.6.1'	1.00	0.67	0.80	3
'3.7.1'	1.00	0.51	0.68	35
'3.7.2'	1.00	0.50	0.67	32
'3.8.1'	0.75	0.32	0.45	102
'3.8.2'	1.00	0.15	0.26	33
'3.9.1'	1.00	0.11	0.20	9
'3.9.2'	0.88	0.42	0.57	36
'3.9.3'	0.00	0.00	0.00	13
'3.a.1'	0.86	0.38	0.52	16
'3.b.1'	0.92	0.28	0.43	39
'3.b.2'	0.75	0.64	0.69	215
'3.b.3'	0.57	0.22	0.31	79
'3.c.1'	0.88	0.32	0.47	47
'3.d.1'	0.93	0.33	0.49	42

- The model is best use for some of the categories:
 - 3.3.1
 - 3.6.1
- Categories that might need other model:
 - 3.1.2
 - 3.9.3
- The model might need to be more tuned for categories:
 - 3.4.1
 - 3.b.2

Alternative Models

KNN Classifier

	precision	recall	f1-score
'3.1.1'	0.57	0.27	0.36
'3.1.2'	0.33	0.17	0.22
'3.2.1'	0.53	0.22	0.31
'3.2.2'	0.29	0.11	0.15
'3.3.1'	0.86	0.55	0.67
'3.3.2'	0.83	0.53	0.65
'3.3.3'	0.81	0.49	0.61
'3.3.4'	1.00	0.12	0.21
'3.3.5'	0.71	0.35	0.47
'3.4.1'	0.76	0.57	0.65
'3.4.2'	0.50	0.36	0.42
'3.5.1'	0.69	0.56	0.62
'3.5.2'	0.75	0.40	0.52
'3.6.1'	0.00	0.00	0.00
'3.7.1'	0.90	0.54	0.68
'3.7.2'	0.90	0.56	0.69
'3.8.1'	0.54	0.28	0.37
'3.8.2'	0.60	0.18	0.28
'3.9.1'	1.00	0.33	0.50
'3.9.2'	0.82	0.50	0.62
'3.9.3'	0.40	0.15	0.22
'3.a.1'	0.82	0.56	0.67
'3.b.1'	0.69	0.28	0.40
'3.b.2'	0.68	0.63	0.66
'3.b.3'	0.44	0.29	0.35
'3.c.1'	0.61	0.30	0.40
'3.d.1'	0.73	0.38	0.50

- KNN Classifier is best for

- 3.3.4
- 3.9.1

- Random Forest is best for:

- 3.1.1
- 3.4.1

- Both models still need tuning for other categories as well:

- KNN have high precision but mid range recall
- Random Forest have high precision but low recall

Both models show promised for categories that were low in the product (ridge) model evaluation score

Random Forest

	precision	recall	f1-score
'3.1.1'	1.00	0.23	0.38
'3.1.2'	0.00	0.00	0.00
'3.2.1'	1.00	0.11	0.20
'3.2.2'	0.50	0.05	0.10
'3.3.1'	0.97	0.38	0.55
'3.3.2'	1.00	0.26	0.41
'3.3.3'	0.94	0.38	0.54
'3.3.4'	0.00	0.00	0.00
'3.3.5'	0.67	0.06	0.11
'3.4.1'	1.00	0.43	0.60
'3.4.2'	0.00	0.00	0.00
'3.5.1'	0.67	0.12	0.21
'3.5.2'	1.00	0.13	0.24
'3.6.1'	1.00	0.33	0.50
'3.7.1'	0.90	0.26	0.40
'3.7.2'	1.00	0.25	0.40
'3.8.1'	0.88	0.14	0.24
'3.8.2'	0.83	0.15	0.26
'3.9.1'	0.00	0.00	0.00
'3.9.2'	0.78	0.19	0.31
'3.9.3'	0.00	0.00	0.00
'3.a.1'	1.00	0.19	0.32
'3.b.1'	1.00	0.18	0.30
'3.b.2'	0.79	0.60	0.68
'3.b.3'	0.58	0.14	0.22
'3.c.1'	0.90	0.19	0.32
'3.d.1'	0.71	0.12	0.20

Recommendation and the Future

Improve by:



Unique Models

Optimized unique models
for each individual
categories

The best text classifier to use:

Ridge Classifier

to help the UN to classify which contender
is relevant to which of the 27 categories



Thank you

Appreciate your time



Appendix

The Categories

- The code and what it is

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