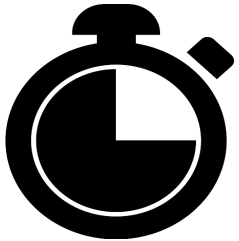
Environmental
Innovations Initiative
University of Pennsylvania

Interdisciplinary Education and Research: Data-informed projects

Project Web Scraper

Project Web Scraper

Problem statement:



Time consuming

High man power

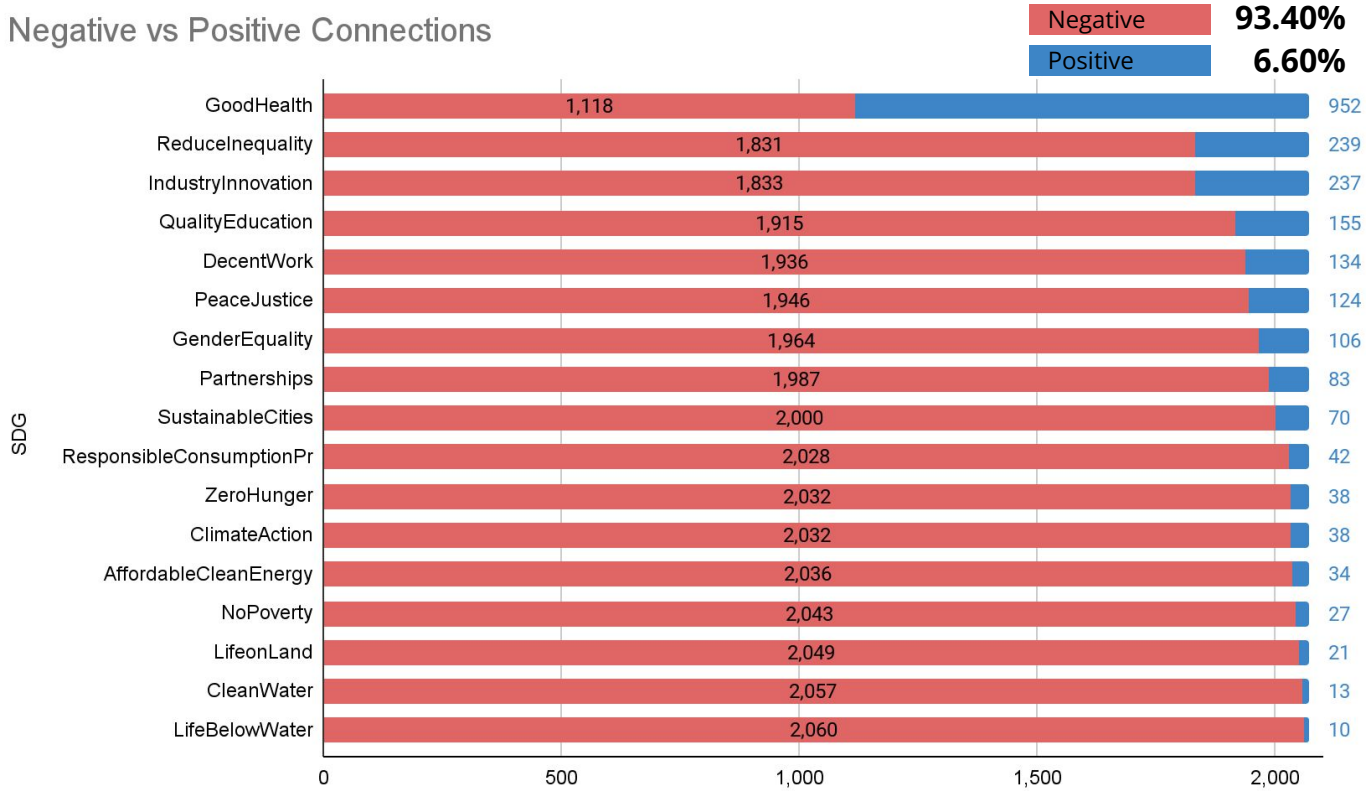
Human error

Objectives:

Develop an end-to-end solution for connecting Penn’s faculty members with the SDGs, from data collection, data storing, and data visualization.

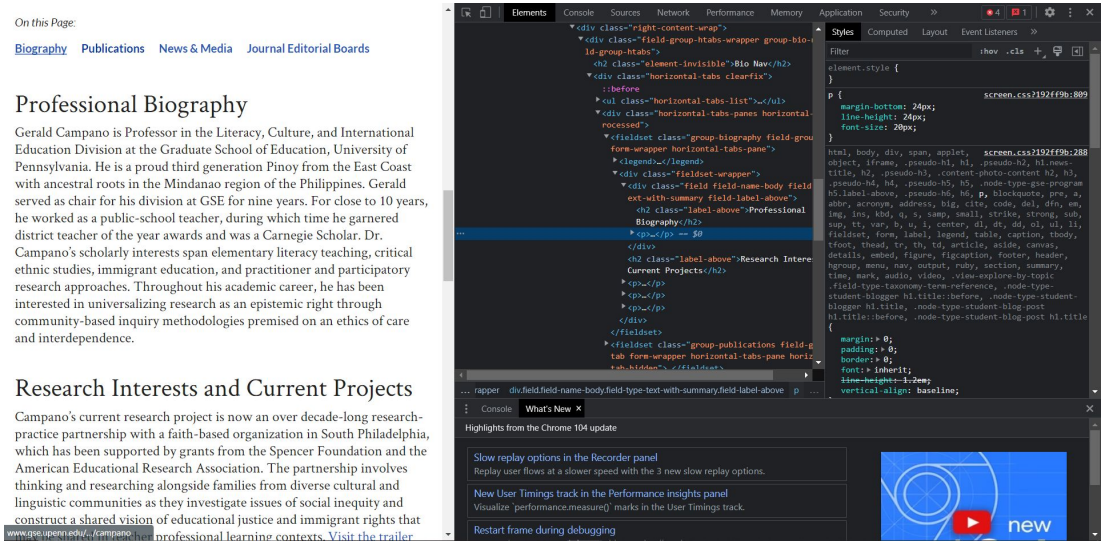
Training Data Exploration - 29/07/2022

The data for this analysis is taken from the 29th of July's database. There are a total of 2,070 rows of data that are used for this analysis. As seen below, the available data is imbalanced. On average, each Penn faculty member only has 1 connection with one of the SDGs.



Web Scraping Process

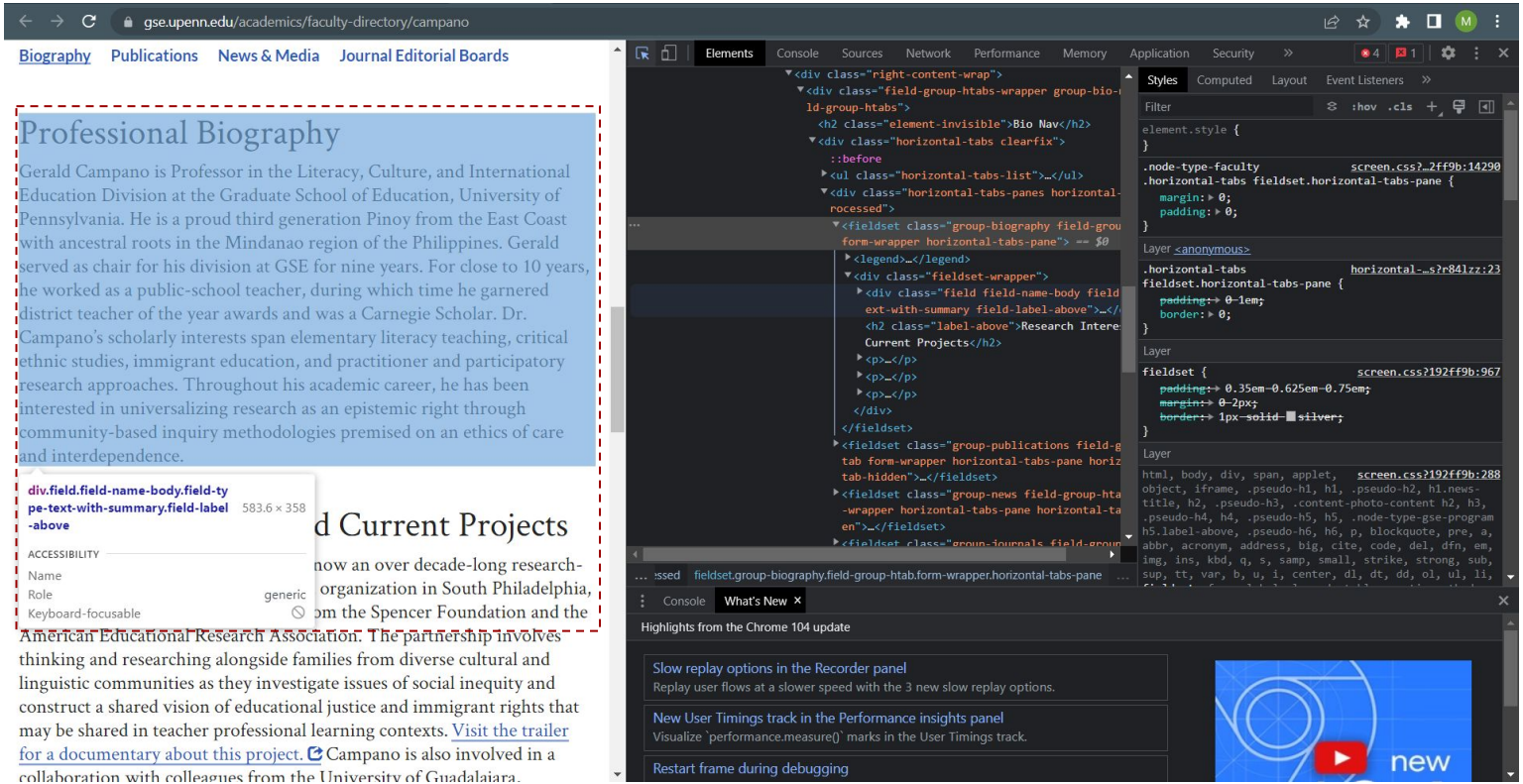
Pass a URL into the web scraper → Extract all p and div tags that contains more than 100 characters → Preprocess the text before fit them into the models



- Lower
- Remove numbers
- Remove punctuation marks
- Remove stopwords
- Lemmatize
- Remove words that have less than 3 characters

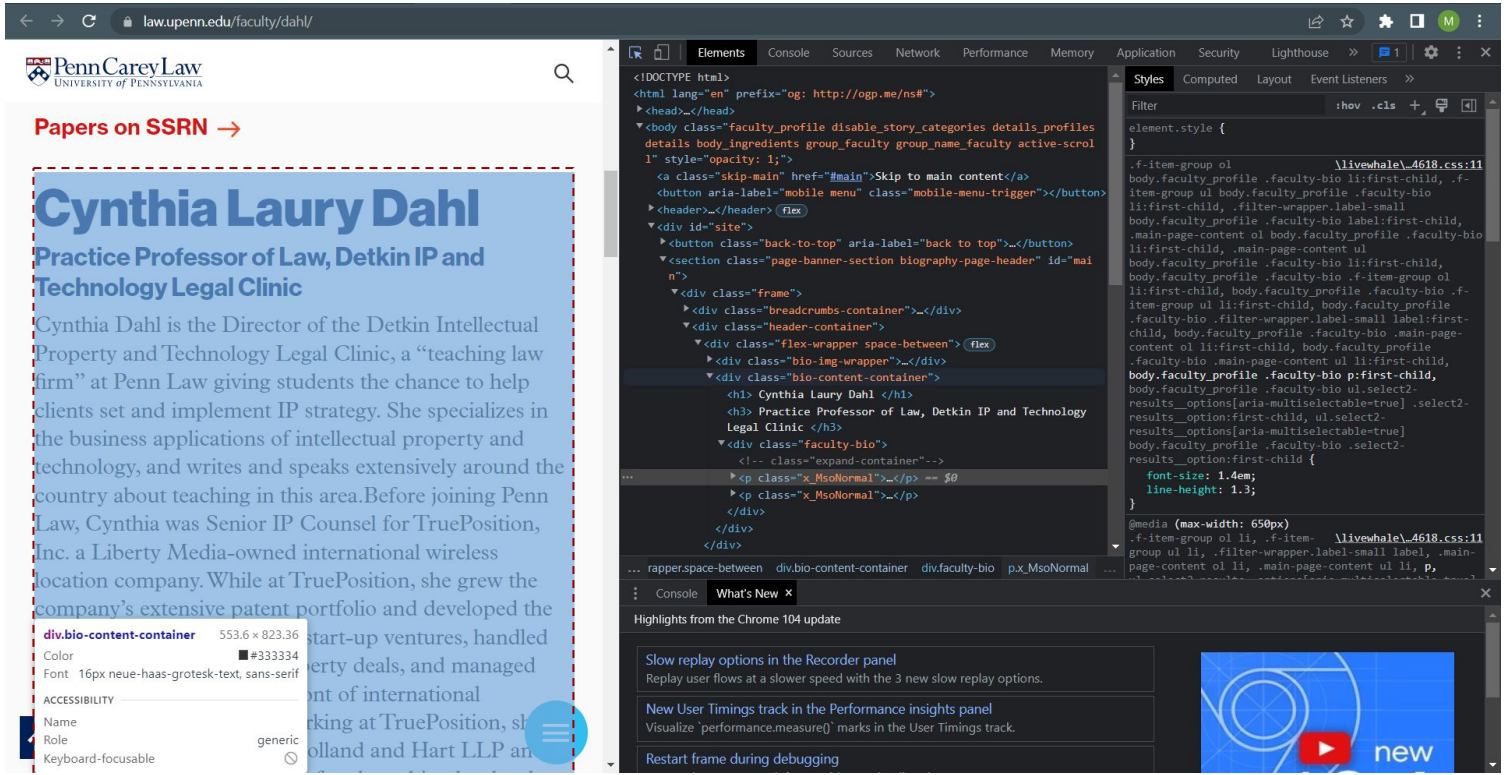
Web Scraping Process - Challenges

In order to extract the correct information, we need to specify the section as div and then navigate to the field class.



Web Scraping Process - Challenges

If we go to a different website, law.upenn.edu, we have to specify div and navigate to bio-content. This means that in order to extract only the relevant information, we need to create a custom scraper for every homepage. Therefore, our current method of scraping all div and p tags, will also download non-relevant information from the websites.



Web Scraping Process

Gerald Campano is Professor in the Literacy, Culture, and International Education Division at the Graduate School of Education, University of Pennsylvania. He is a proud third generation Pinoy from the East Coast with ancestral roots in the Mindanao region of the Philippines. Gerald served as chair for his division at GSE for nine years. For close to 10 years, he worked as a public-school teacher, during which time he garnered district teacher of the year awards and was a Carnegie Scholar. Dr. Campano's scholarly interests span elementary literacy teaching, critical ethnic studies, immigrant education, and practitioner and participatory research approaches. Throughout his academic career, he has been interested in universalizing research as an epistemic right through community-based inquiry methodologies premised on an ethics of care and interdependence.

campano professor literacy culture international education division graduate school education university pennsylvania proud third generation pinoy east coast ancestral root mindanao region philippine gerald served chair division gse nine year close year worked public school teacher time garnered district teacher year award carnegie scholar campano scholarly interest span elementary literacy teaching critical ethnic study immigrant education practitioner participatory research approach throughout academic career interested universalizing research epistemic right community based inquiry methodology premised ethic care interdependence

Web Scraping Process

Remove non SDG related words → Texts are ready to analyze

List of SDG keywords:

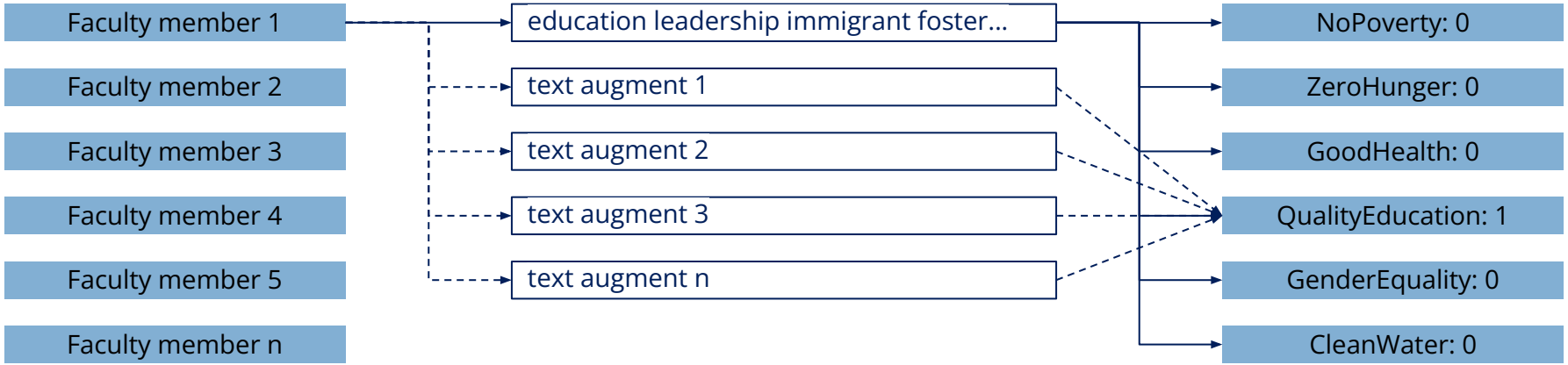
- abortion
- abuse
- access
- accident
- acidic
- acidification
- acidity
- addict
- adolescence
- advocates
- afforest
- ... and 544 others

education leadership immigrant foster work culture education education work foster youth voice work immigrant leader young immigrant partnership infrastructure partnership cultural immigrant refugee city cooperate access culture education identity leadership entrepreneurship immigrant refugee culture...

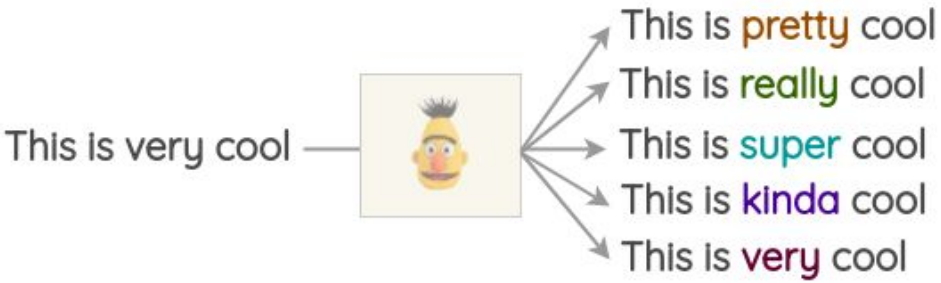
*Keywords are taken from Aurora, Leicester, and SDSN Australia.
*The full list of keywords can be found here.

Data Augmentation

One of the ways to work with imbalanced data is to augment the minority class using the Bert model. The model will create new data that have a similar meaning to the original data.



Bert data augmentation example:



TF-IDF (Term Frequency-Inverse Document Frequency)

A statistical measure that evaluates how relevant a word is to a document in a collection of documents. The higher a word repeats and the more that word exists in a corpus, the lower the correlation between the word to the SDG.

Mathematical equation

$$TF = \frac{\text{No. of repetition of words in document}}{\text{No. of words in document}}$$

$$IDF = \log\left(\frac{\text{No. of documents}}{\text{No. of documents containing words}}\right)$$

$$TF\text{-}IDF = TF * IDF$$

Model Selection - Training & Validation phase

We transform the data into multi dependent variable (multi-label classification) instead of a multi-class dependent variable. We then fit the data to predict each SDG category.

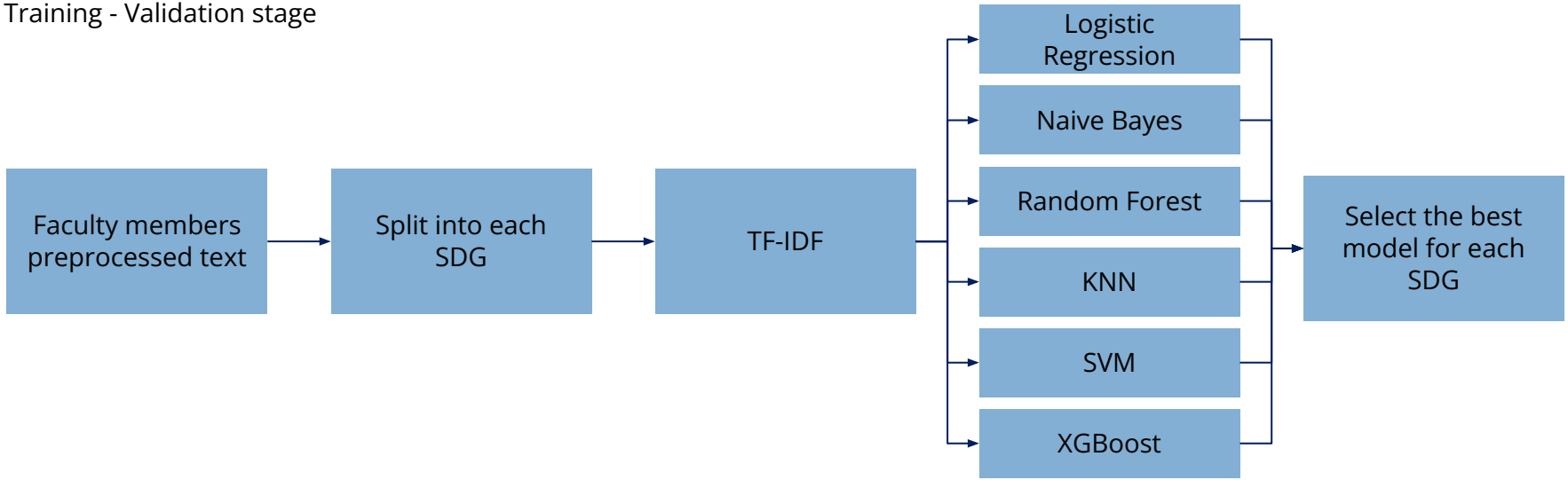
Multi label classification

PennID	NoPoverty	ZeroHunger	GoodHealth
101	1	0	1

Multi class classification

PennID	SDG
101	NoPoverty; GoodHealth

Training - Validation stage



Model Selection - Metrics Evaluation

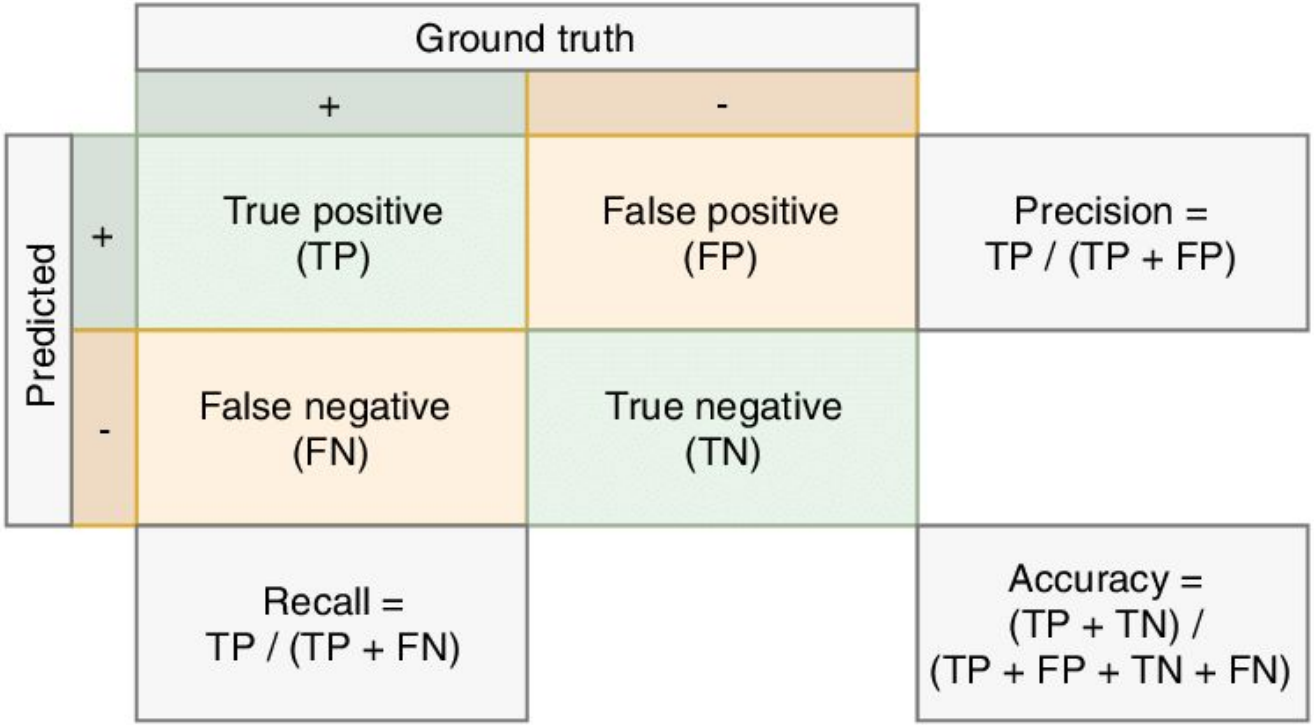
Since our data is extremely imbalance, accuracy will tell a biased result. Even if the model predicts everything as 0, it will still produce a high accuracy. Therefore, We will focus on the F1 score, with a preference for higher recall relative to precision.

Metrics	Description	Formula
Accuracy	Overall effectiveness of a classifier	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Class agreement of the data labels with positive labels given by the classifier	$\frac{TP}{TP+FP}$
Recall/Sensitivity	Effectiveness of a classifier to identify positive labels	$\frac{TP}{TP+FN}$
F1 score	Combination of precision and sensitivity	$2\frac{PRC \cdot SNS}{PRC+SNS}$

*TP (True positives): predicted positive, actual positive
*TN (True negatives): predicted negative, actual negative
*FP (False positives): predicted positive, actual negative
*FN (False negatives): predicted negative, actual positive

Model Selection - Confusion Matrix

For reference



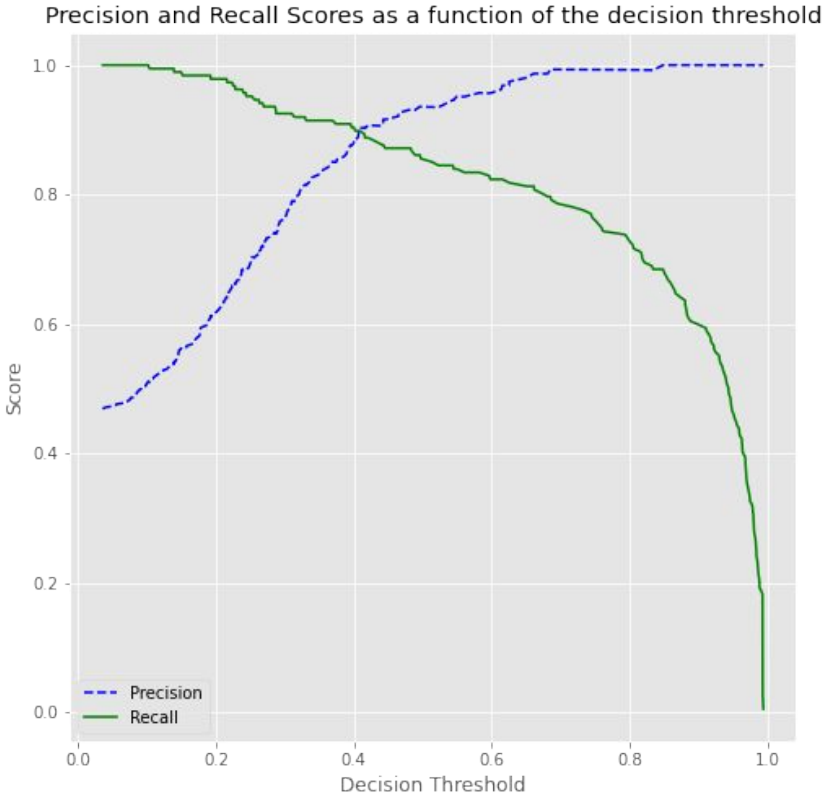
Model Selection - Training Data Performance

The data shown are only for the positive class. The average **precision, recall, and F1 scores for the positive class are 0.475, 0.509, and 0.437**. In comparison, the average scores for the negative class are 0.973, 0.964, and 0.97.

SDG	Precision	Recall	F1 Score	Support	Threshold
NoPoverty	0.5	0.4	0.44	5	0.43
ZeroHunger	0.33	0.43	0.38	7	0.55
GoodHealth	0.94	0.86	0.9	187	0.41
QualityEducation	0.81	0.45	0.58	29	0.29
GenderEquality	0.39	0.43	0.41	21	0.5
CleanWater	0.6	1	0.75	3	0.95
AffordableCleanEnergy	1	0.14	0.25	7	0.29
DecentWork	0.33	0.76	0.46	25	0.65
IndustryInnovation	0.61	0.39	0.47	44	0.26
ReduceInequality	0.42	0.67	0.52	46	0.5
SustainableCities	0.21	0.33	0.26	12	0.7
ResponsibleConsumptionProduction	0.29	0.75	0.41	8	0.6
ClimateAction	0.12	0.43	0.19	7	0.59
LifeBelowWater	0.05	0.5	0.08	2	0.65
LifeonLand	0.67	0.5	0.57	4	0.47
PeaceJustice	0.45	0.42	0.43	24	0.225
Partnerships	0.33	0.31	0.32	16	0.25

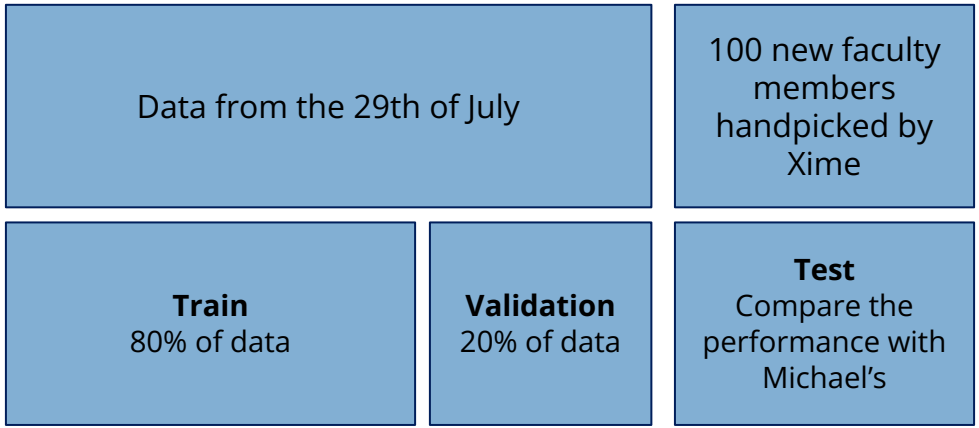
Model Selection - Precision, Recall, and Threshold

The threshold numbers defined for each of the model are based on the prioritization of recall instead of precision. In other words, it is better to have more false positives than false negatives.



Model Selection - Test phase

To further validate the model performance, we selected 100 new data consisting of faculty members from different schools. We then measure the accuracy of this phase by comparing the model's prediction with Michael's analysis.



Test Data Exploration - 100 Random URLs

To test the model, we use 100 new URLs that the model has not seen yet. We then compare the result with a manual analysis run by Michael to measure the accuracy.

Model Selection - Test Data Performance

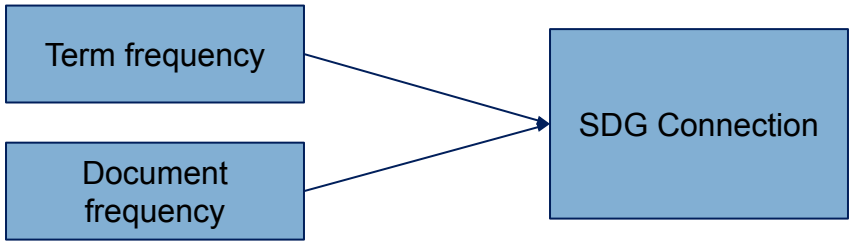
The average **precision, recall, and F1 scores for the positive class are 0.483, 0.366, and 0.379**. In comparison, the average scores for the negative class are 0.966, 0.981, and 0.969.

SDG	Precision	Recall	F1 Score	Support	Threshold
NoPoverty	0	0	0	3	0.43
ZeroHunger	0.33	0.5	0.4	2	0.55
GoodHealth	0.95	0.99	0.97	76	0.41
QualityEducation	1	0.08	0.14	26	0.29
GenderEquality	0	0	0	0	0.5
CleanWater	1	1	1	1	0.95
AffordableCleanEnergy	1	0.33	0.5	3	0.29
DecentWork	0.67	0.4	0.5	5	0.65
IndustryInnovation	1	0.5	0.67	14	0.26
ReduceInequality	0.58	0.78	0.67	9	0.5
SustainableCities	0.5	0.5	0.5	2	0.7
ResponsibleConsumptionProduction	0	0	0	4	0.6
ClimateAction	0.67	1	0.8	2	0.59
LifeBelowWater	0	0	0	0	0.65
LifeonLand	1	0.5	0.67	2	0.47
PeaceJustice	0	0	0	5	0.225
Partnerships	0	0	0	0	0.25

Model Selection - Test Model Analysis

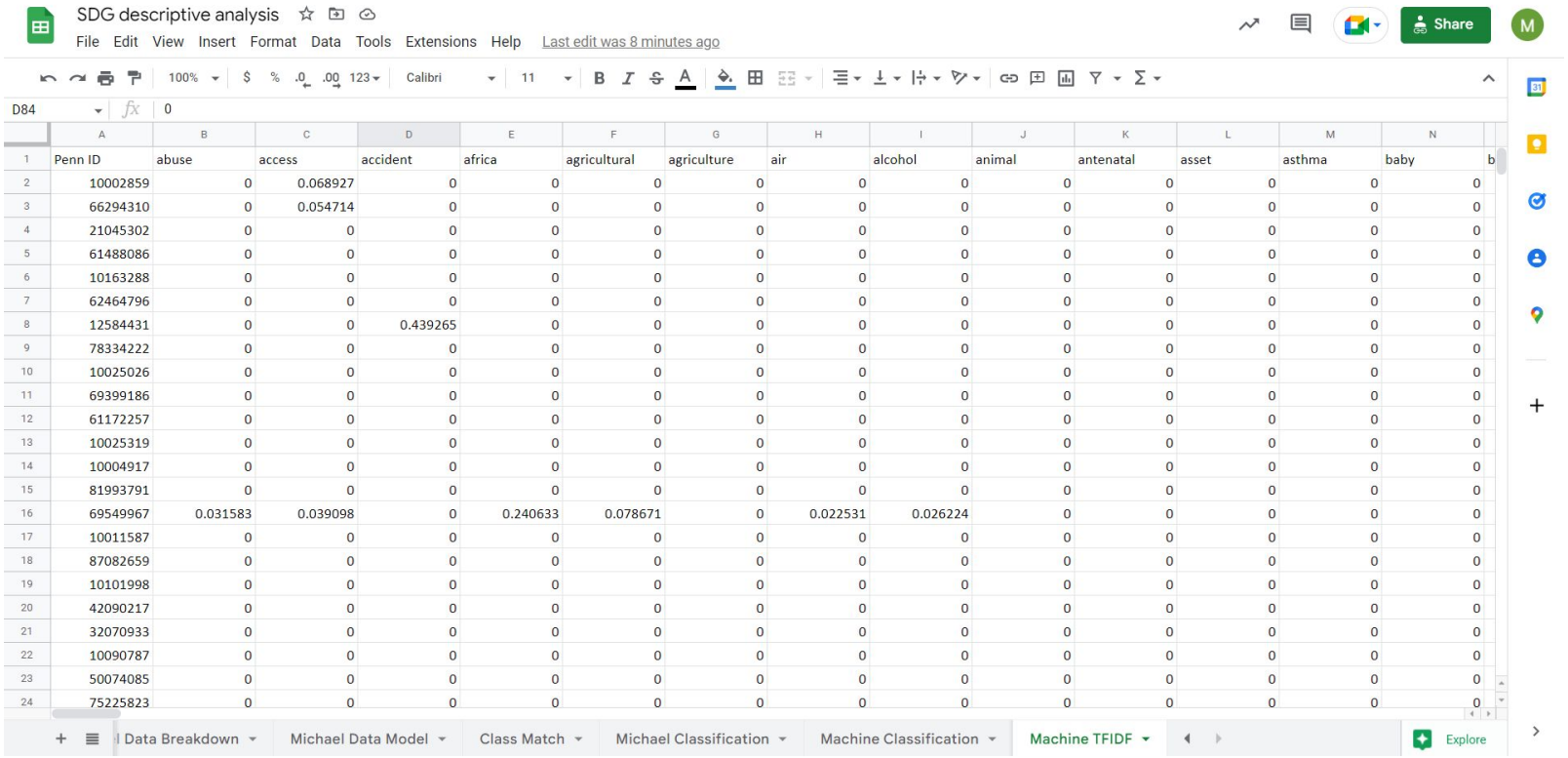
There are 2 reasons why the model has a lower score relative to the training-validation score:

1. **Incomplete keywords dictionary**
    a. To offset the amount of text/information scraped from a given website, we remove non-specific keywords that have a slightly neutral meaning and often appear across documents.
    b. Keywords such as professor, college, program, study, and student are purposely removed from the list of keywords.
    c. However, there are other keywords that are not listed yet from our three main sources, Aurora, Leicester, and SDSN Australia, such as neurochemistry, neuroscience, pathologies, etc.
    d. In total, there are 62 unique keywords that Michael identify (taken only from the false negative) to connect with the SDGs, while the keywords can only match 14 of them.
2. **Low correlation between TF-IDF score on the matched keywords**
    a. Most of the matched keywords, such as economic, appear across different documents but are not always identified with any SDG, making the word have a low correlation with the SDG.
    b. On the other hand, a keyword such as health also appears a lot, but most documents with the word health in them are linked to the GoodHealth SDG.



Model Selection - How the data looks like

Each word will have a score between 0 - 1, where the value represents the level of distinctness and importance of a given word to a document. Currently, not all keywords related to the SDGs are used. This is because the web scraper is still scraping a lot of non-relevant information from each faculty member's website.



SDG descriptive analysis

File Edit View Insert Format Data Tools Extensions Help Last edit was 8 minutes ago

Share

D84 0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Penn ID	abuse	access	accident	africa	agricultural	agriculture	air	alcohol	animal	antenatal	asset	asthma	baby
2	10002859	0	0.068927	0	0	0	0	0	0	0	0	0	0	0
3	66294310	0	0.054714	0	0	0	0	0	0	0	0	0	0	0
4	21045302	0	0	0	0	0	0	0	0	0	0	0	0	0
5	61488086	0	0	0	0	0	0	0	0	0	0	0	0	0
6	10163288	0	0	0	0	0	0	0	0	0	0	0	0	0
7	62464796	0	0	0	0	0	0	0	0	0	0	0	0	0
8	12584431	0	0	0.439265	0	0	0	0	0	0	0	0	0	0
9	78334222	0	0	0	0	0	0	0	0	0	0	0	0	0
10	10025026	0	0	0	0	0	0	0	0	0	0	0	0	0
11	69399186	0	0	0	0	0	0	0	0	0	0	0	0	0
12	61172257	0	0	0	0	0	0	0	0	0	0	0	0	0
13	10025319	0	0	0	0	0	0	0	0	0	0	0	0	0
14	10004917	0	0	0	0	0	0	0	0	0	0	0	0	0
15	81993791	0	0	0	0	0	0	0	0	0	0	0	0	0
16	69549967	0.031583	0.039098	0	0.240633	0.078671	0	0.022531	0.026224	0	0	0	0	0
17	10011587	0	0	0	0	0	0	0	0	0	0	0	0	0
18	87082659	0	0	0	0	0	0	0	0	0	0	0	0	0
19	10101998	0	0	0	0	0	0	0	0	0	0	0	0	0
20	42090217	0	0	0	0	0	0	0	0	0	0	0	0	0
21	32070933	0	0	0	0	0	0	0	0	0	0	0	0	0
22	10090787	0	0	0	0	0	0	0	0	0	0	0	0	0
23	50074085	0	0	0	0	0	0	0	0	0	0	0	0	0
24	75225823	0	0	0	0	0	0	0	0	0	0	0	0	0

Data Breakdown | Michael Data Model | Class Match | Michael Classification | Machine Classification | Machine TFIDF | Explore

Model Selection - Summary and Recommendations

There are three reasons why the model has such low performance:

1. The current web scraper is still scraping non-relevant text from a given URL
    - This is due to the different HTML and tags used to store the main text/information
    - The current web scraper is scraping all texts under p and div tags that contain more than 100 characters
    - Additionally, the current web scraper is not able to extract data from javascript websites
2. Several keywords are not included in the current SDG keywords dictionary
    - Keywords such as professor and teaching are such common words that can reduce the accuracy of the models
3. Data is highly imbalanced

Recommendations

The recommendations to improve the model performance:

1. Customized the web scraper to extract the exact information based on the number of unique URL
    - Find the correct tags/sections for each of the 267 unique URL/homepage that contains the accurate information
2. Expand the keyword dictionary once the web scraper is able to extract the right information
    - Run a keyword analysis to determine the balance between relevance and occurence
3. Run a cross study with either Aurora or OSDG
    - Replicate Aurora and OSDG's studies and uses their positive data to expand our training database