Which Sustainable Development Goals (SDG) are linked in the country's National Determined Contributions (NDC) document?

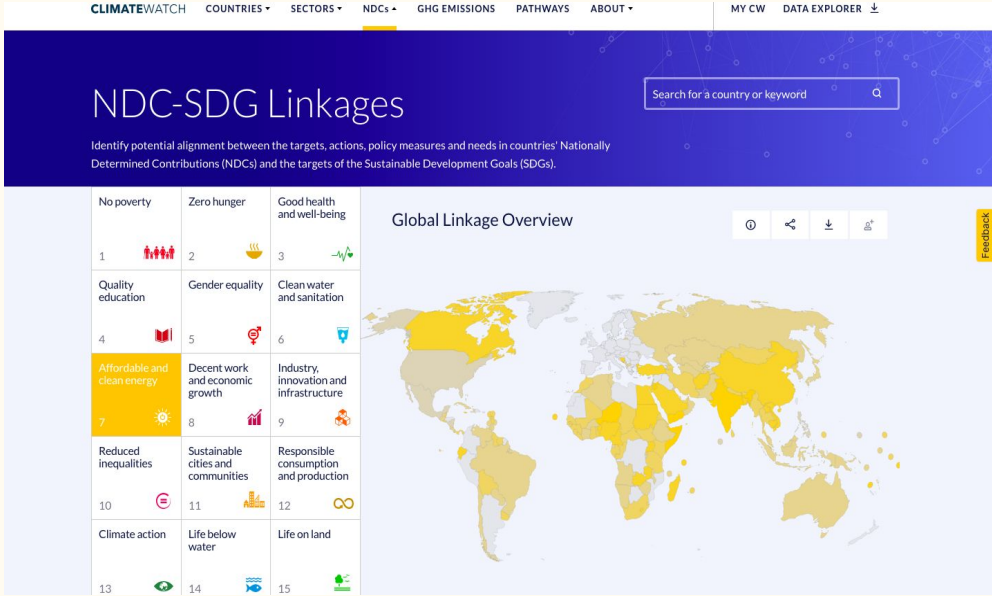
—

Text analysis, Classification modeling

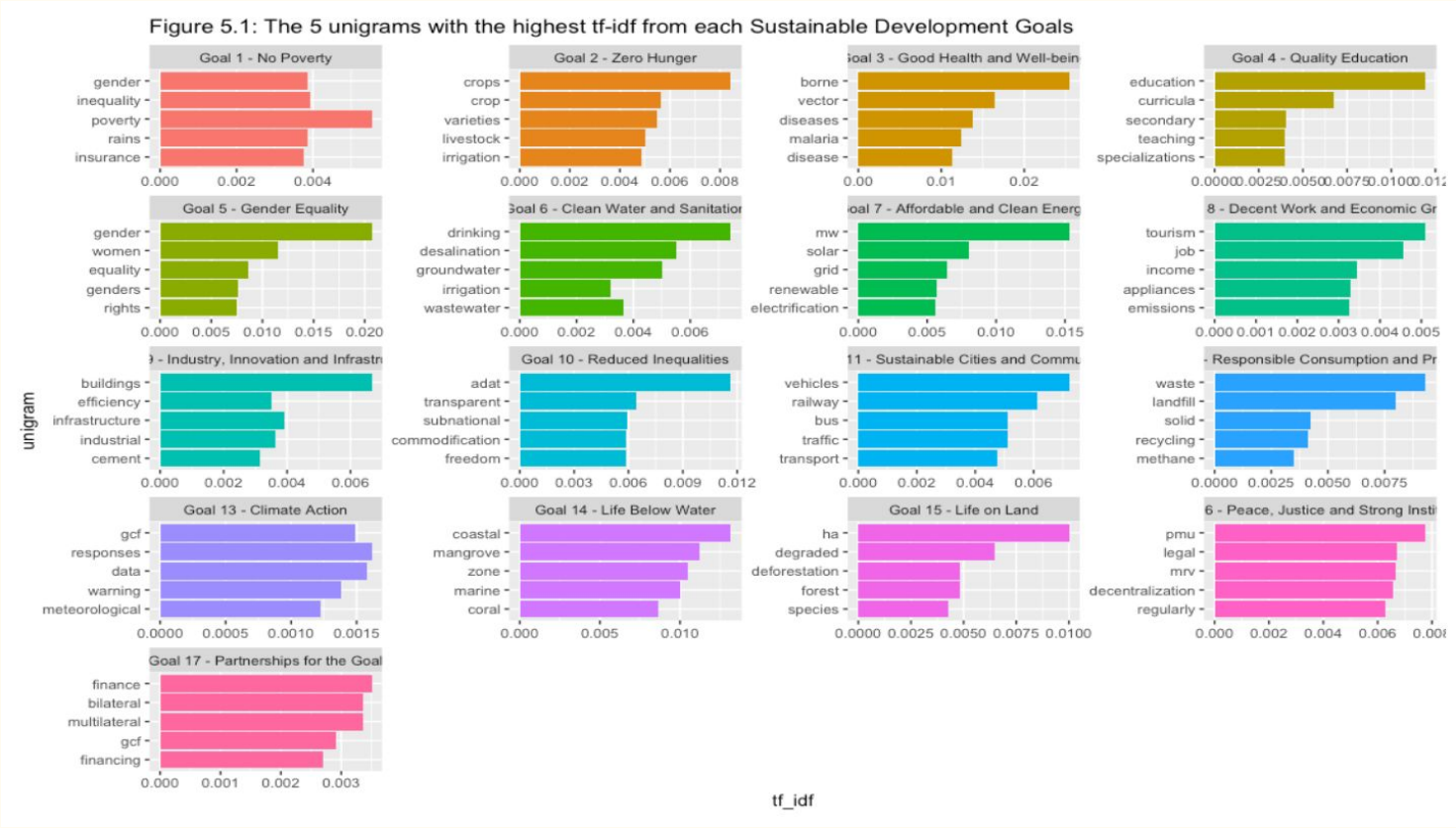
Shiying Wang

Overview

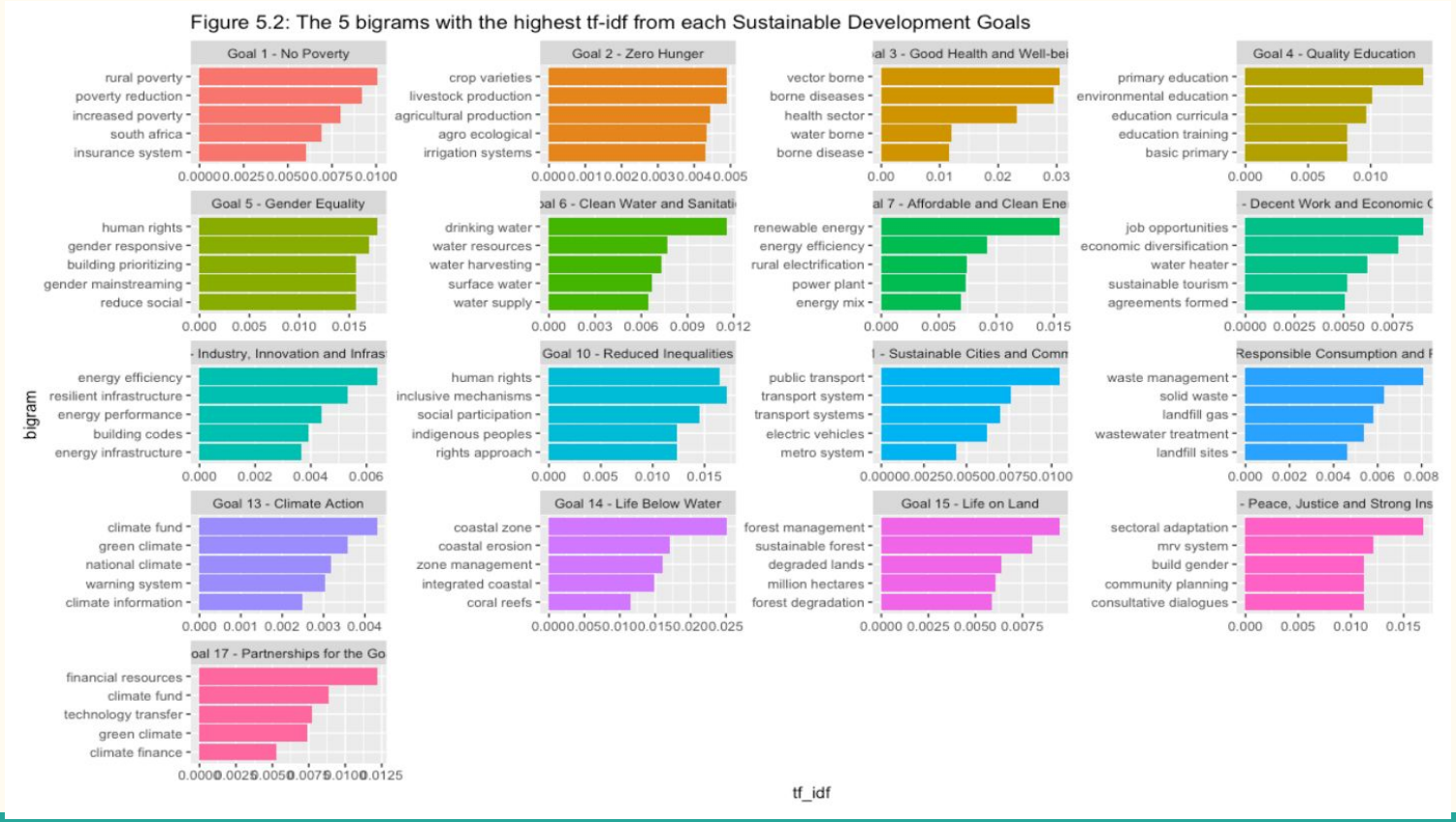
- Motivation
- Raw data
  - 9602 observations
    - id: unique id for each sentence
    - ndc_text: sentence from countries's NDC document
    - goal: tagged SDG
- Methods
  - n_gram tokenization, TF-IDF, Document Term Matrix, Classification Modeling (Random forest, KNN, Neural network)
- Result
  - EDA
  - Random forest with 55% accuracy



Top 5 unigram with the highest tf-idf in each SDG



Top 5 bigram with the highest tf-idf in each SDG



Classification Modeling Result

9602 observations

231 predictors

17 categories

8 models

Future work: trends by countries, political regions, better sampling method, using bigrams/trigrams in the models...

Model <fctr>	Accuracy <fctr>
Random Forest Model_50	54.8%
Random Forest Model_100	55.2%
KNN_80	36.8%
KNN_100	37.3%
KNN_300	42.3%
Neural Network_3	51.4%
Neural Network_4	53.4%
Neural Network_5	54.4%

Some takeaways..

- Not only the balance of variance and bias, but sometimes also a balance of modeling accuracy and computational costs.
- “Don’t get obsessed with doing anything perfectly, but instead build one reasonably well working pipeline to get a sense of the process and learn about the data. ”
- Yihui Xie’s blog https://yihui.org/cn/recipe/