Predicting Sustainable Development Indices from Geolocated Text

Cara Van Uden [*1] **Andrew Gaut** [*1] **Won Kyung Do** [*2] **Aman Bansal** [*1]

Abstract

We leverage readily-available natural language data, scraped from Wikipedia, to predict localized indices (asset, sanitation, women's education) relevant to the UN's Sustainability Goals. We explore the impact of different text embedding extraction methods and model architectures on performance in this small data task. We explore logistic regression models, feedforward DNNs, and NLP-CNNs. We use geolocated and extracted "relevant" sentence embeddings to achieve ROC-AUC scores of 0.80 (logistic regression model), 0.70 (logistic regression model), and 0.81 (feedforward DNN model) for asset, sanitation, and women's education index classification, respectively.

1. Motivation

In 2015, the United Nations (UN) released 17 Sustainable Development Goals (SDGs) aimed at promoting global peace and prosperity (UN, 2015). Prior work has successfully applied machine learning – typically computer vision – to promote progress towards these goals (Yeh et al., 2020). Such tasks take satellite images of an area as input and attempt to predict metrics related to prosperity for that area, such as the level of wealth. Such predictions have been used in Bangladesh, Togo, and Uganda to intelligently decide where and how to send aid (Yeh et al., 2021; Nature, 2020). Recently, a dataset called SustainBench with labels for many such tasks was released (Yeh et al., 2021).

In our work, we explore the efficacy of employing text features – as opposed to image features – for some of the SustainBench tasks. As the internet provides text data with rich information pertinent to many SustainBench tasks (like news articles mentioning the economic prosperity of a region, for instance), it could be a very useful data source. While prior work has explored using images and text jointly as features for some SustainBench tasks (Uzkent et al., 2019; Sheehan et al., 2019), to our knowledge we are the first to explore using text alone. In particular, we use scraped geolocated Wikipedia data to predict the asset index, sanitation index, and average women's educational achievement level for the location associated with the geolocated Wikipedia article.

2. Related Work

As SustainBench is a brand new dataset, no studies yet exist which attempt the tasks defined by the dataset. However, a wealth of studies worked on similar datasets and achieved promising results using satellite imagery to predict wealth levels and other items relevant to the SDGs like crop yield and crop classification (Jean et al., 2016; Yeh et al., 2020; Steele et al., 2017; Holloway et al., 2018; Lee et al., 2021; M Rustowicz et al., 2019; Wang et al., 2018). Our work aims to achieve comparable performance to prior studies in the task of poverty prediction with only text-based data to prove the efficacy of using text-based data alone.

Our work also borrows from well-developed sub-fields of Natural Language Processing (NLP). Specifically, predefined word lists have been used to flag text sources for hate speech and to detect gender bias and dehumanization in embedding spaces (Xiang et al., 2012; Williams & Burnap, 2016; Nobata et al., 2016; Dinakar et al., 2012; Bolukbasi et al., 2016; Mendelsohn et al., 2020). We borrow this tactic and create pre-defined lists of words relevant to each of the targets we consider – asset index, women's educational attainment, sanitation index, and clean water index – to select relevant sentences to feed to our sentence encoder to create the features we use for our model.

Prior work has also explored the usage of CNNs for classification problems with text features (Kim, 2014; Bitvai & Cohn, 2015a). Most relevantly to our work, prior work has employed CNNs in a very small data text-based problem – only 1,000 training documents – and achieved very high performance (Bitvai & Cohn, 2015b). Since our dataset is small, this inspired us to utilize a CNN for this project.

[*]Equal contribution [1]Department of Computer Science, Stanford University, Stanford, CA [2]Department of Mechanical Engineering, Stanford University, Stanford, CA. Correspondence to: Cara Van Uden <cvanuden@stanford.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

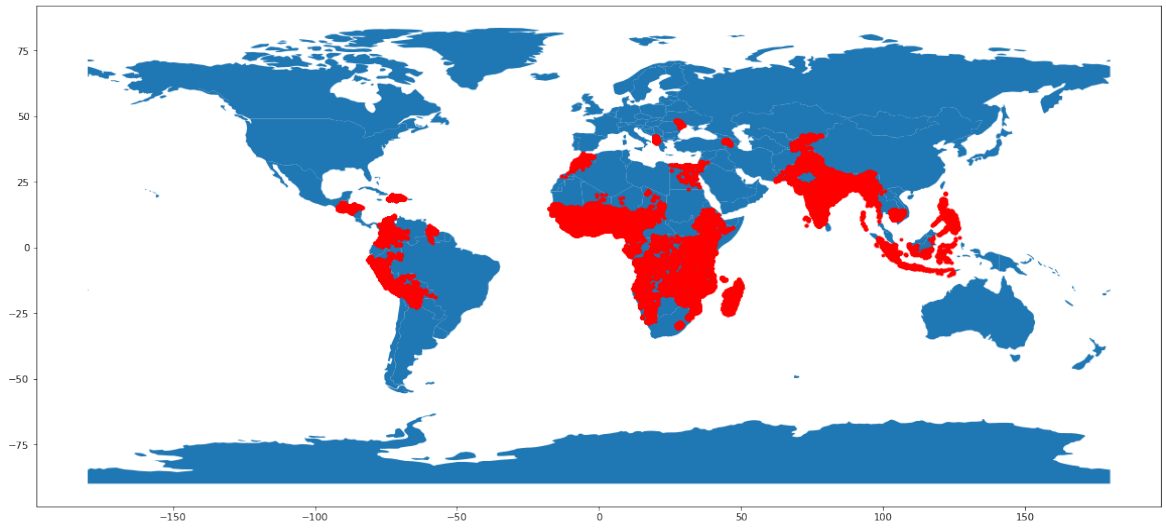

Figure 1. Cluster locations included in SustainBench.

3. Dataset

3.1. SustainBench

SustainBench (Yeh et al., 2021) is a collection of 15 benchmark tasks across 7 UN Sustainable Development Goals (SGDs), including tasks related to economic development, agriculture, health, education, water and sanitation, climate action, and life on land. In our analysis, we focus on four of these tasks - predicting 1) asset index, a measure of economic development, 2) sanitation index, a measure of sanitation quality, and 4) women education measured in years, a measure of social equality and education. SustainBench obtains their labels from Demographic and Health Surveys (DHS), which surveys residents in developing countries (ICF, 1996-2019).

Based on our exploratory data analysis, we found that SustainBench features such as child mortality (for which 71.4% of labels were either NULL or 0) and women's BMI (which had vastly different distributions and correlations with the other features between countries) may not be the best choices for our analysis. Instead, we chose to focus on three features - asset index, sanitation index, and women's education index - in our downstream analyses. (Please see our exploratory data analysis notebook (`notebooks/eda.ipynb`) for more details.)

3.1.1. Asset Index

The asset index used by SustainBench represents the aggregate level of wealth in a given area centered at some location specified by a (latitude, longitude) pair. In particular, the asset index of a household over an asset set $\{1, ..., k\}$ is $A = \sum_{j=1}^{k} a_j b_j$, where a_j is an indicator varaible indicating the household posesses item j and b_j is a weighting factor indicating the importance of that asset. The maximum likelihood estimators of the b_j values are the principle components of the data (Moser & Felton, 2007; Filmer & Scott, 2012). SustainBench's asset index is the first principle component.

The authors of SustainBench drew from DHS surveys conducted between 1996 and 2019 to assemble asset wealth data for 2,079,036 households living in 86,936 clusters across 48 countries. In our exploratory data analysis, we found that there were no asset indices available before 2004. We additionally found that asset index ranges between -3.8 and 3.6 and has a bimodal distribution, with peaks below and above the median around 0. Across countries and years, asset index is also highly correlated with the sanitation and water indices we predict.

3.1.2. Sanitation Index

Our sanitation index prediction task involves predicting a cluster-level sanitation index. These cluster-level indices represent the "static" water or sanitation quality of a cluster at a given point in time. Again, the authors of SustainBench drew from DHS surveys conducted between 1996 and 2019 to assemble data for 2,143,329 households living in 89,271 clusters across 49 countries. Note that the sanitation indices were readily available as scalar values in the DHS surveys, unlike asset index, so no further preprocessing was required.

In our exploratory data analysis, we found that there were no sanitation indices available before 2004. We additionally found that sanitation index ranges between 1 and 5 and has two peaks at 1 (3.51% of the labels are exactly 1) and 5 (10.22% of the labels are exactly 5). Across countries

Index	Features	Precision	Recall	F-1 Score	Accuracy	ROC-AUC
asset index	sentence	0.83 (0.88/0.70)	0.85 (0.85/0.75)	0.82 (0.87/0.73)	0.82	0.80
asset index	document	0.61 (0.66/0.54)	0.63 (0.85/0.27)	0.60 (0.74/0.36)	0.63	0.56
asset index	sentence, document	0.79 (0.86/0.64)	0.78 (0.82/0.70)	0.79 (0.84/0.67)	0.78	0.76
sanitation index	sentence	0.73 (0.66/0.80)	0.70 (0.87/0.53)	0.70 (0.75/0.64)	0.70	0.70
sanitation index	document	0.60 (0.65/0.54)	0.61 (0.75/0.42)	0.60 (0.69/0.47)	0.61	0.59
sanitation index	sentence, document	0.73 (0.66/0.79)	0.71 (0.86/0.56)	0.70 (0.75/0.65)	0.71	0.71
women's education index	sentence	0.75 (0.72/0.79)	0.74 (0.88/0.57)	0.73 (0.79/0.67)	0.74	0.73
women's education index	document	0.63 (0.68/0.52)	0.66 (0.91/0.19)	0.61 (0.78/0.28)	0.66	0.55
women's education index	sentence, document	0.65 (0.62/0.69)	0.63 (0.88/0.33)	0.60 (0.73/0.45)	0.63	0.61

Table 1. Precision, recall, F1 score, accuracy, and ROC-AUC for logistic regression models trained on each feature and predicting each target. Note that Precision, Recall, and F-1 Score values are given with the weighted average as (weighted average, class 0/class 1). (So, for example, for the asset index with sentence features, the logistic regression model gets 0.83 weighted average precision, 0.88 precision on class 0, and 0.70 precision on class 1.)

and years, sanitation index was highly correlated with asset index.

3.1.3. Women's Education

Our women's education prediction task involves predicting a cluster-level women's education index. This cluster-level index represents average years of educational attainment by women of reproductive age (15-49) of a cluster at a given point in time. The authors of SustainBench drew from DHS surveys conducted between 1996 and 2019 to assemble data for 3,013,286 women living in 122,435 clusters across 56 countries. Note that the women's education index was readily available as scalar values in the DHS surveys, unlike asset index, so no further preprocessing was required.

In our exploratory data analysis, we found that women's education index ranges between 0 and 15 and has a nearly normal distribution, except for a large peak at 0 (1.17% of the labels are exactly 0). Aggregated countries and years, unlike the other indices we predict, women's education is not consistently correlated with other indices such as asset index. However, depending on country women's education can be highly correlated with asset index.

3.2. Wikipedia

To obtain text data from which we could extract features, we needed to find text pertinent to at least one location for which SustainBench provides labels. Since some Wikipedia articles are tagged with geolocations (which indicate the article is written about that geolocation) and since Wikipedia articles about locations typically contain information relevant to our prediction targets, we chose to scrape data from Wikipedia. We scraped all Wikpedia articles with a tagged geolocation and later matched them to locations which have labels in the SustainBench dataset (see Section 4.1).

Note that we also tried scraping data from the Global Dataset of Events, Language and, Tone (GDELT) and Twitter (Leetaru & Schrodt, 2013). However, in both cases, text was written in a plethora of languages, with some of these languages being very low resource languages, and so it was difficult to find ways to encode these data sources. In the case of Twitter, we also needed approval to use the Academic Research API level to be able to search for tweets by geolocation, and our request was not approved in time for us to scrape the data. However, we have functioning code for scraping data from both sources and have included these files in our submitted code repository.

4. Data Preprocessing

4.1. Mapping Wikipedia Articles to Cluster Locations

Once we obtained the geolocated Wikipedia articles, we had to match the article locations with the locations corresponding to the labels from the SustainBench dataset. For each Wikipedia article, we found the SustainBench label with the closest location to that Wikipedia article. If the closest label had a (latitude, longitude) pair within a radius of 3 of the (latitude, longitude) pair of the article, then we assigned that article to that SustainBench label. Otherwise, we concluded the article had no matching label and so did not use it.

4.2. Extracting Relevant Keywords

Given text data from geolocated Wikipedia articles, we wanted to extract the most relevant sentences for each target index. We chose to use gensim's pre-trained FastText model as our word embedder because it was trained on Wikipedia and news data, two data sources that we either used downstream in the current analysis or plan to use in future work (Mikolov et al., 2018).

To generate a dictionary of relevant keywords for each target index, we first generated a dictionary of starter relevant keywords for each target index - for example, the starter keyword seeds for asset index were `["wealth", "poverty", "money", "income", "inequality"]`. Then, we gathered the 25 closest keywords in word embedding space for each of the these starter keyword seeds, as well as the 25 closest keywords in word embedding space to the entire list of starter keyword seeds. We deduplicated these new keywords and used these as our relevant keywords for the index. This resulted in 142 keywords related to asset index, 189 keywords related to sanitation index, and 139 keywords related to women's education index. We experimented with different values for number of neighbor keywords to fetch, but settled on 25 neighbor keywords per seed because we wanted 100-200 keywords per index, and found that we started getting many duplicated keywords with values greater than 25.

4.3. Extracting Relevant Sentences

For each target index and each Wikipedia article, we extracted all of the sentences with at least one keyword in the target index's relevant keyword dictionary. We embedded all of these relevant sentences and used them in our downstream basic regression and classification models.

For the NLP-CNN specifically, we wanted to find the most relevant sentences out of this subset of relevant sentences. For each of these sentences, we also extracted a "relevance score" of how many times a keyword appeared in the sentence. To find the most relevant sentences, we extracted all sentences in an article with the maximum relevance score. We used these most relevant sentences as input to our NLP-CNN. An example relevant sentence for asset index, with relevance score 2, is `This was necessary in order to maintain control over spiraling borrowing costs, since the relatively poor emirate finances investments by raising funds on financial markets.`

4.4. Sentence Embedding

For each target index and each Wikipedia article, we now had a set of relevant sentences. We embedded all of these sentences using our chosen pre-trained sentence embedder - huggingface's `all-MiniLM-L6-v2` sentence transformer model (hug). This pre-trained model maps sentences and paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search. Next, we averaged together all sentence embeddings for each cluster location and target index - note that a survey location could have multiple Wikipedia articles mapped to it, each of which could have one or more relevant sentences. In this way, we had a single averaged sentence embedding for each each cluster location and each target index.

4.5. Document Embedding

In addition to extracting and embedding relevant sentences only, we also tried embedding entire Wikipedia articles. To do this, we trained a Doc2Vec model on the collection of geolocated Wikipedia articles which had already been matched to at least one corresponding DHS label (as described in Section 4.1) (Le & Mikolov, 2014). We configured the model to create embeddings of dimension 300 and trained the model for 10 epochs.

5. Model

5.1. Baseline Models

5.1.1. Regression

For each target index and set of features of sentence embeddings, document embeddings, and concatenated sentence and document embeddings, we trained kernel ridge regression models to predict the target index using the features. We performed hyperparameter optimization over kernel type (linear or RBF) and alpha (50 values log-spaced between 0.01 and 1000) using grid search with successive halving.

5.1.2. Classification

Next, we binarized each of our target indices with thresholds based on our insights from our exploratory data analysis.

Our threshold for asset index was 0, threshold for sanitation index was 3, and threshold for women's education index was 5. If we had an imbalanced dataset - the minority class made up less than 30% of the data - we resampled from each clas. We oversampled from the minority class using SMOTE (Chawla et al., 2002), and undersampled from the majority class using Tomek Links (Elhassan & Aljurf, 2016). For each target index and set of features of sentence embeddings, document embeddings, and concatenated sentence and document embeddings, we trained logistic regression models to predict the target index class using the scaled features. We performed hyperparameter optimization over values of C (20 values log-spaced between 0.1 and 100) using grid search with successive halving.

5.2. Feedforward DNN

We performed hyperparameter optimization using the ray tune library (Liaw et al., 2018). We generated 4 sample neural networks for each target index and set of features and performed random hyperparameter search over learning rate (log uniform between 0.001 and 0.1) and hidden layer size (choice between 128, 256, and half of the input size hidden layers). Each sample neural network had a single hidden layer, used cross-entropy loss as its loss function, used Adam as its optimization algorithm (Kingma & Ba, 2014), and was trained for 10 epochs. As for our other classification models, we used the same threshold for classifying each index (asset index: 0, sanitation index: 3, and women education index: 5).

5.3. NLP-CNN

As discussed in Section 2, CNNs are sometimes used for sentence classification in NLP and have been shown to be effective even when little training data is available (Kim, 2014; Bitvai & Cohn, 2015a). In particular, in this architecture, sentences are fed to the CNN and a word embedding model converts each word in the sentence into its corresponding embedding and these embeddings are stacked into a tensor in the order the words appear in the sentence.

Based on results from our exploratory data analysis, we truncated the sentence data to have a maximum sentence length of 1000. We have 4291 data examples for each index. Because we have so little data, we use a pre-trained word embedding model; we use the FastText model specifically (Joulin et al., 2016; Mikolov et al., 2018). Our CNN has three convolutional layers with filter height/width of 3, 4, and 5 respectively and with 100 filters for each layer. Then, we pass the output through a single fully connected layer with dropout and softmax the output to obtain the distribution over classes. To inspect the performance of pre-trained model, we trained the model in three ways: 1) model with randomized initial value, 2) pre-trained model with freezing the embedding layer, and 3) pre-trained model with fine-tuning the embedding layers.

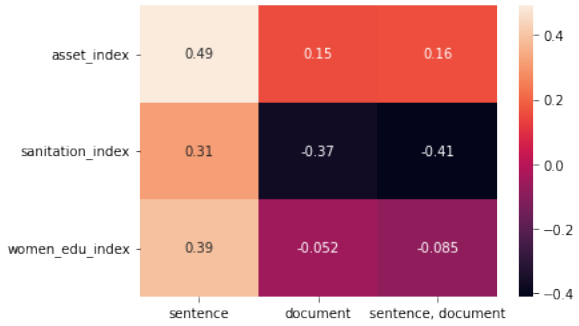

Figure 2. Kernel ridge regression performance (R^2) across features and target indices.

As for our other classification models, we used the same threshold for classifying each index (asset index: 0, sanitation index: 3, and women education index: 5).

6. Results

6.1. Baseline Models

6.1.1. Regression

We achieved an R^2 of 0.49 for asset index prediction, 0.31 for sanitation index prediction, and 0.39 for women's education index prediction, all using our relevant/target sentence embeddings. See Figure 2 for results for all feature input types. Notably, we did not see improved performance when using document embeddings or using both relevant sentence and document embeddings.

6.1.2. Classification

We achieved an ROC-AUC of 0.80 for asset index prediction, 0.70 for sanitation index prediction, and 0.73 for women's education index prediction, all using our relevant/target sentence embeddings. See Figure 4 and Table 3.1.1 for detailed results for all feature input types. Notably, we did not see improved performance when using document embeddings or using both relevant sentence and document embeddings.

6.2. Feedforward DNN

We achieved an ROC-AUC of 0.75 for asset index prediction, 0.66 for sanitation index prediction, and 0.81 for women's education index prediction, all using our relevant/target sentence embeddings. See Table 6.2 for detailed results for all feature input types. Notably, we did not see improved performance when using document embeddings or using both relevant sentence and document embeddings. We also saw improved performance (over the basic logistic regression

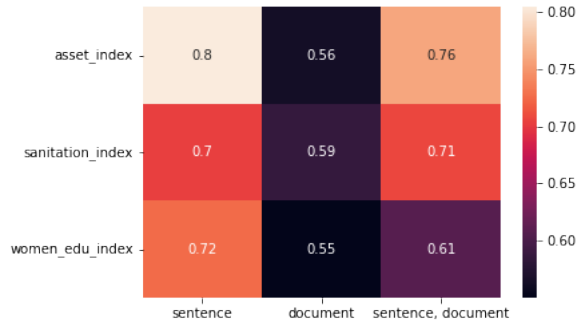

Figure 3. Logistic regression performance (ROC-AUC) across features and target indices. See table 3.1.1 for more details, including precision, recall, F1 score, accuracy, and ROC-AUC.

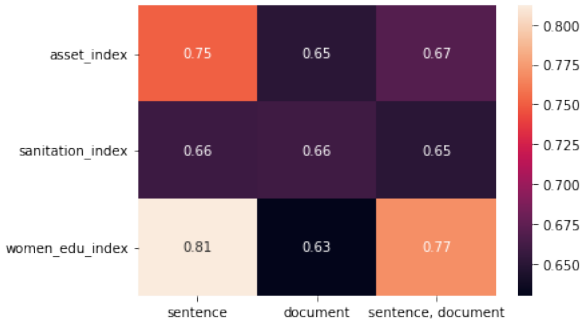

Figure 4. Feedforward DNN performance (ROC-AUC) across features and target indices. See table 3.1.1 for more details, including precision, recall, F1 score, accuracy, and ROC-AUC.

models) only for predicting women's education index.

6.3. NLP-CNN

The result for each index using NLP-CNN model is shown in Table 6.3. Table 6.3 shows the best accuracy among 20 epochs with batch size 50. Overall, we see that the models mostly classify all examples as belonging to only one class - note the precision, recall, and F1 scores per class, as well as the difference between accuracy and ROC-AUC across models. We expected to see that the pre-trained model outperforms the randomly initialized model. All indices except asset index shows that the pre-trained model with fine-tuning the embedding layer has the best performance. We can expect that the model for asset index has been overfitted during using pre-trained model since the result has similar accuracy but slightly worse ($1, 2\%$ difference).

7. Discussion

Our results demonstrate that it is possible to solely use text-based data to classify location clusters by various indices (asset, sanitation, and women's education) relevant

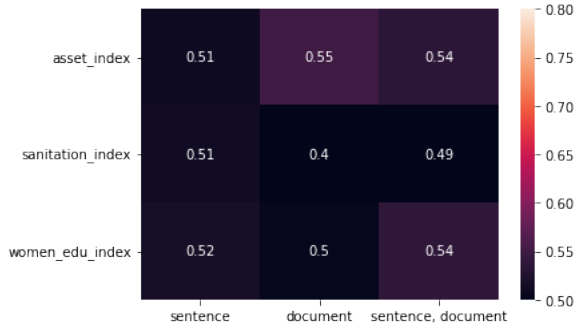

Figure 5. NLP-CNN performance (ROC-AUC) across features and target indices. See table 3.1.1 for more details, including precision, recall, F1 score, accuracy, and ROC-AUC.

to the UN SGDs. Moreover, our results show that small, non-resource intensive models like logistic regression can perform well at this task. Moreover, obtaining sentence embeddings requires relatively little computational power as well, since one need only use a pre-trained model. Because the method is relatively scalable, real-world institutions will be able to create models similar to ours by surveying the populations in a collection of locations, finding text written about those locations, embedding that text, and then training small models on that data.

However, the text data used in this study is largely static and updated infrequently, which is not ideal because there may be a large time gap between the improvement in area's wealth index or sanitation index and the corresponding edits to the area's Wikipedia article reflecting these changes. We thus recommend that future work utilize more frequently updated text sources with geolocation data. We recommend that future work specifically look to use the Global Dataset of Events, Language, and Tones (GDELT) or Twitter data (Dredze et al., 2016). GDELT is a massive database updated every 15 minutes containing information about recently written news articles – including the (latitude, longitude) values of locations mentioned in the article, among other things. Future work can query the database for (latitude, longitude) values matching or nearby locations which have SustainBench labels, then use the corresponding articles in the database as input data for that label. Similarly, the Twitter API allows for searching for tweets by the geolocation of the device they were tweeted from; future work can find tweets with geolocations close to locations with SustainBench labels and use those as inputs for those labels. Note that using GDELT or Twitter also makes it significantly easier to use our method to deploy models to new locations and to get real-time data, since data for a variety of locations is more readily found from these two data sources than from Wikipedia. Such data would let us make more fine-grained real-time predictions, a massive improvement over

Index	Features	Precision	Recall	F-1 Score	Accuracy	ROC-AUC
asset index	sentence	0.78 (0.86/0.59)	0.76 (0.77/0.73)	0.76 (0.81/0.65)	0.76	0.75
asset index	document	0.67 (0.74/0.55)	0.66 (0.69/0.62)	0.67 (0.72/0.58)	0.66	0.65
asset index	sentence, document	0.72 (0.84/0.46)	0.65 (0.60/0.74)	0.66 (0.70/0.57)	0.65	0.67
sanitation index	sentence	0.66 (0.64/0.69)	0.66 (0.76/0.56)	0.66 (0.69/0.62)	0.66	0.66
sanitation index	document	0.67 (0.72/0.59)	0.67 (0.71/0.61)	0.67 (0.71/0.60)	0.67	0.66
sanitation index	sentence, document	0.65 (0.66/0.64)	0.65 (0.63/0.66)	0.65 (0.65/0.65)	0.65	0.65
women's education index	sentence	0.82 (0.90/0.73)	0.80 (0.72/0.90)	0.80 (0.80/0.80)	0.80	0.81
women's education index	document	0.66 (0.76/0.48)	0.63 (0.64/0.62)	0.64 (0.70/0.54)	0.63	0.63
women's education index	sentence, document	0.77 (0.74/0.80)	0.77 (0.76/0.78)	0.77 (0.73/0.45)	0.77	0.77

Table 2. Precision, recall, F1 score, accuracy, and ROC-AUC for feedforward DNN models trained on each feature and predicting each target. Note that Precision, Recall, and F-1 Score values are given with the weighted average as (weighted average, class 0/class 1).

the infrequently collected on-the-ground surveys currently used.

If we had had more computational and financial resources, we could have scraped more data from GDELT. We had used Google BigQuery to extract information about relevant articles, but such queries were costly due to the size of the GDELT dataset and, moreover, GDELT only provides the URL to the article in question, and so we had to then scrape articles after each query, which was time intensive due to the low bandwidth of the AWS servers we used to scrape the data. With more data and resources, we could also have trained much deeper models.

References

Huggingface sentence transformers. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2021-11-25.

Bitvai, Z. and Cohn, T. Non-linear text regression with a deep convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 180–185, Beijing, China, July 2015a. Association for Computational Linguistics. doi: 10.3115/v1/P15-2030. URL https://aclanthology.org/P15-2030.

Bitvai, Z. and Cohn, T. Non-linear text regression with a deep convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 180–185, 2015b.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30, 2012.

Dredze, M., Osborne, M., and Kambadur, P. Geolocation for twitter: Timing matters. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1064–1069, 2016.

Index	Model	Precision	Recall	F-1 Score	Accuracy	ROC-AUC
asset index	Random initial value	0.71 (0.80/0.33)	0.79 (0.98/0.05)	0.72 (0.88/0.08)	0.79	0.510
asset index	Pre-trained with frozen layer	0.69 (0.80/0.29)	0.79 (0.98/0.03)	0.71 (0.88/0.06)	0.79	0.547
asset index	Pre-trained with fine-tuned layer	0.69 (0.80/0.25)	0.78 (0.96/0.05)	0.71 (0.87/0.08)	0.78	0.535
sanitation index	Random initial value	0.72 (0.79/0.44)	0.78 (0.98/0.06)	0.71 (0.88/0.11)	0.78	0.514
sanitation index	Pre-trained with frozen layer	0.67 (0.79/0.22)	0.77 (0.97/0.03)	0.70 (0.87/0.05)	0.77	0.397
sanitation index	Pre-trained with fine-tuned layer	0.83 (0.79/0.99)	0.79 (0.99/0.02)	0.70 (0.88/0.03)	0.79	0.493
women's education index	Random initial value	0.66 (0.76/0.36)	0.74 (0.97/0.05)	0.66 (0.85/0.09)	0.74	0.518
women's education index	Pre-trained with frozen layer	0.69 (0.76/0.47)	0.75 (0.96/0.11)	0.68 (0.85/0.17)	0.75	0.504
women's education index	Pre-trained with fine-tuned layer	0.65 (0.76/0.33)	0.72 (0.92/0.12)	0.67 (0.83/0.17)	0.72	0.537

Table 3. Precision, recall, F1 score, accuracy, and ROC-AUC for NLP-CNN models trained on each models.

Elhassan, T. and Aljurf, M. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. *Global J Technol Optim S*, 1, 2016.

Filmer, D. and Scott, K. Assessing asset indices. *Demography*, 49(1):359–392, 2012.

Holloway, J., Mengersen, K., and Helmstedt, K. Spatial and machine learning methods of satellite imagery analysis for sustainable development goals. In *Proceedings of the 16th Conference of International Association for Official Statistics (IAOS)*, pp. 1–14. International Association for Official Statistics (IAOS), 2018.

ICF. Demographic and health surveys (various), 1996-2019.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

Kim, Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181.

Kingma and Ba. Adam: A method for stochastic optimization. 2014. URL https://arxiv.org/abs/1412.6980.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. PMLR, 2014.

Lee, J., Brooks, N. R., Tajwar, F., Burke, M., Ermon, S., Lobell, D. B., Biswas, D., and Luby, S. P. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118 (17), 2021.

Leetaru, K. and Schrodt, P. A. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pp. 1–49. Citeseer, 2013.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

M Rustowicz, R., Cheong, R., Wang, L., Ermon, S., Burke, M., and Lobell, D. Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 75–82, 2019.

Mendelsohn, J., Tsvetkov, Y., and Jurafsky, D. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55, 2020.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Moser, C. and Felton, A. The construction of an asset index measuring asset accumulation in ecuador cprc. Technical report, Vol. Working Paper 87). Washington DC: The Brookings Institution, 2007.

Nature. Machine learning can help get covid-19 aid to those who need it most. 2020. URL https://www.nature.com/articles/d41586-020-01393-7.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pp. 145–153, 2016.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D., and Ermon, S. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2698–2706, 2019.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., De Montjoye, Y.-A., Iqbal, A. M., et al. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690, 2017.

UN. Transforming our world: The 2030 agenda for sustainable development. 2015. URL https://sustainabledevelopment.un.org/post2015/transformingourworld/publication.

Uzkent, B., Sheehan, E., Meng, C., Tang, Z., Burke, M., Lobell, D., and Ermon, S. Learning to interpret satellite images using wikipedia. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Wang, A. X., Tran, C., Desai, N., Lobell, D., and Ermon, S. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 1–5, 2018.

Williams, M. L. and Burnap, P. Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211–238, 2016.

Xiang, G., Fan, B., Wang, L., Hong, J., and Rose, C. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1980–1984, 2012.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11 (1):1–11, 2020.

Yeh, C., Meng, C., Wang, S., Driscoll, A., Rozi, E., Liu, P., Lee, J., Burke, M., Lobell, D. B., and Ermon, S. SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 12 2021. URL https://openreview.net/forum?id=5HR3vCylqD.