natureresearch
https://doi.org/10.1038/s42256-020-00235-5

Supplementary information

Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning

In the format provided by the authors and unedited

Supplementary data

We refer to the features of the SVM-KNN-Shocastic model in the "Interventions" section of the paper. The following is an overview of the models, features and process to achieve intervention labeling.

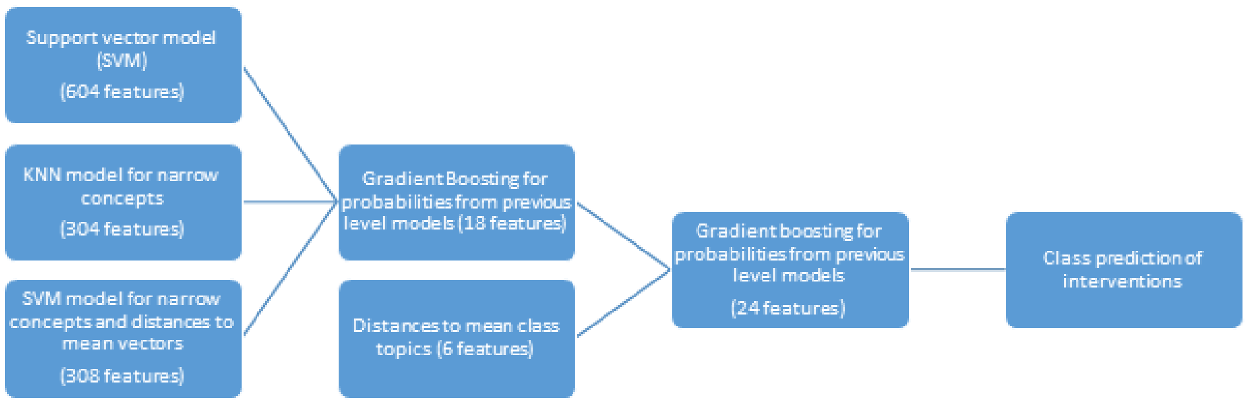

Figure 1 Intervention labeling models

We present the confusion matrix for the intervention labeling task.

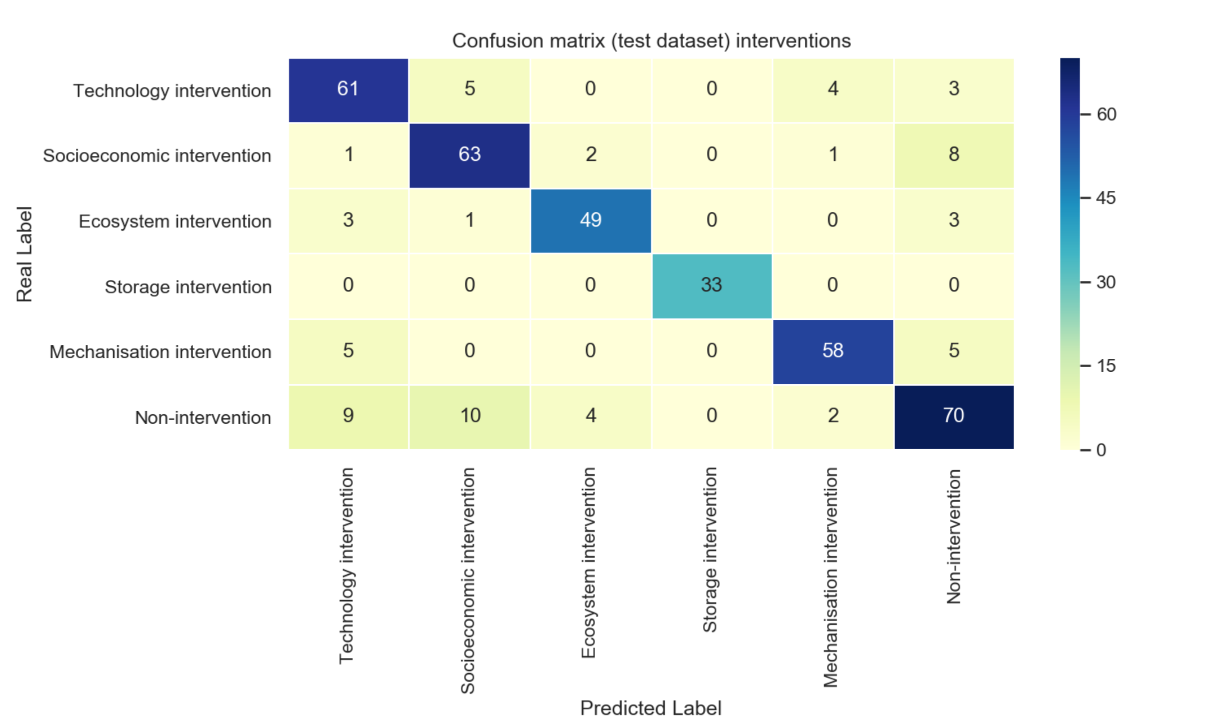

Figure 2: Confusion matrix for the tasks "intervention labeling"

Table 1: Precision, Recall and F1-Measures for interventions labeling			
Intervention class	Precision	Recall	F1-measure
Technology intervention	0.77	0.83	0.79
Socioeconomic intervention	0.8	0.84	0.82
Ecosystem intervention	0.89	0.87	0.88
Storage intervention	1	1	1
Mechanisation intervention	0.89	0.85	0.87
Non-intervention	0.78	0.73	0.76
Average	0.85	0.85	0.85

Table 1 Precision, Recall and F1-Measures for interventions labeling

Additional information about training to label for study design type

The BERT model was trained to predict one of the following study design types: observational study, modelling study, laboratory study, review paper, field study and no category. The training dataset (14,715 articles) was gathered via implicit labelling by searching study type keywords in the articles and, in addition to this data, 1000 articles were labelled manually. Test dataset consists of 250 manually labelled articles. Training the model just on the manually labelled data brings 0.62 F1-measure for the test data, whereas model training on the composed train dataset gives 0.71 F1-measure for the test data. This may be due to overfitting on a small dataset, and bringing more data to the model even with probably incorrect labels increases generality of the model.

Several BERT models (base BERT, Roberta, Albert, SciBERT) were checked for model quality with the settings: 128, 256, 384 and 512 tokens. The SciBERT with 256 tokens gave the best result with F1-measure = 0.76. In this setting the model uses only the first 256 tokens, but to take more information from abstracts, we trained the model on the last 256 tokens too. As far as 93% of articles' abstracts consist of less than 512 tokens, the model will not lose very much if we restrict training only on the first and last 256 tokens. The model performance if we take only the last 256 tokens equals 0.72 F1-measure.

We combined SciBERT on the last hidden layer outputs to train the final model, which will predict a study design type for an article. The input to the final model consists of 3 parts. The first part is a 768 vector, representing the last hidden layer output for the first 256 tokens. The third part is a 768 last hidden layer output for the last 256 tokens. The second part captures the features from the whole abstract: by window sliding with the size 256 tokens and the step 128 tokens, we encode each 256 tokens into a 768 dimensional vector, and after encoding all the parts till the end of the abstract we perform coordinatewise max pooling. The result is a 768 vector which is used as a second part of the input. These three concatenated parts are the input to the 2 layer feedforward neural network, while training dropout rate 0.2 was used to prevent overfitting. The models with 1, 2, 3 layers were checked, but 2 layer neural network gave the best result F1-measure 0.8. The confusion matrix, precision, accuracy, and F1-measure for study design types is shown. The confusion matrix for study design type based on 250 examples.

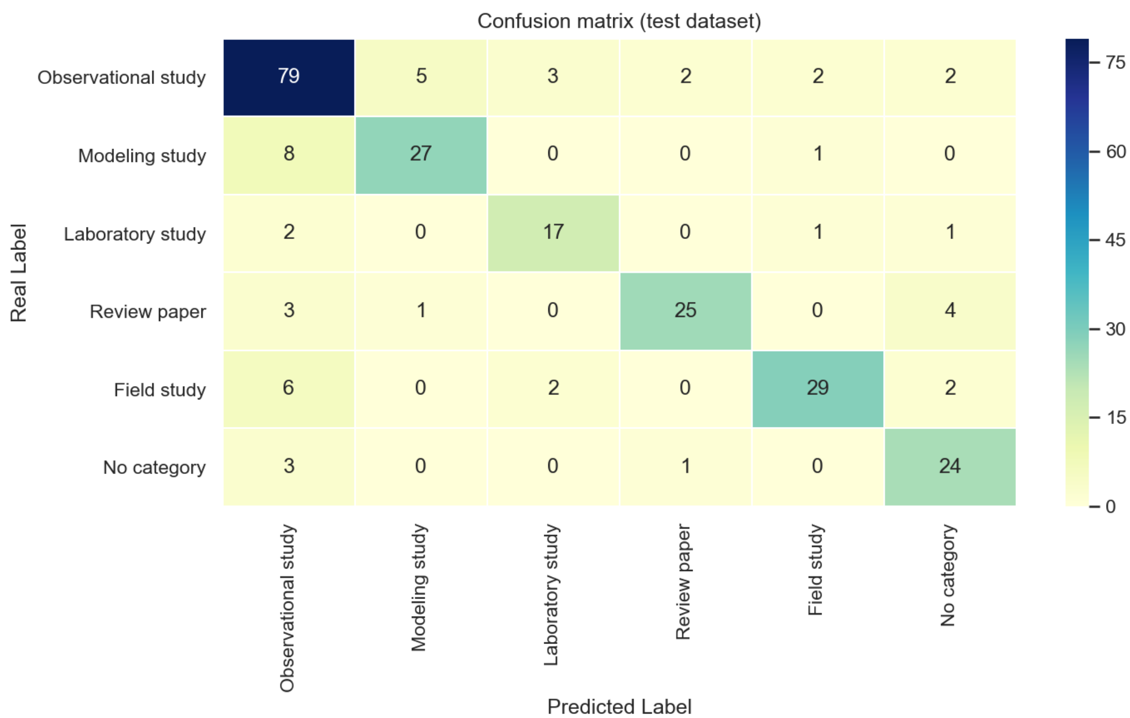

Figure 3 A confusion matrix of the classification task for study design type

Table 2 Precision, Recall and F1-Measures for study design type

Table 2: Precision, Recall and F1-Measures for study design type classification			
Study type	Precision	Recall	F1-measure
Observational study	0.78	0.85	0.81
Modelling study	0.82	0.75	0.78

Laboratory study	0.77	0.81	0.79
Review paper	0.89	0.76	0.82
Field study	0.88	0.74	0.81
No category	0.73	0.86	0.79
Average	0.81	0.79	0.8

Measurements model
Base BERT model was trained on the corpus 5100 examples. Each example represents a sentence with a captured intervention (pretrained using our intervention ontology) or an agricultural crop and numerical data. The model should predict whether a particular numerical data is connected semantically with the found concept in the context of the sentence. The BERT model allows to split the input into two segments and find connections between them, so the first segment is a sentence and the second segment contains a number and an intervention or crop, separated by a space. The model is trained on 128 tokens, as far as neither of examples exceeds this token number. The F1-measure is 0.8 for this model. Confusion matrix and classification reports are shown below.
.

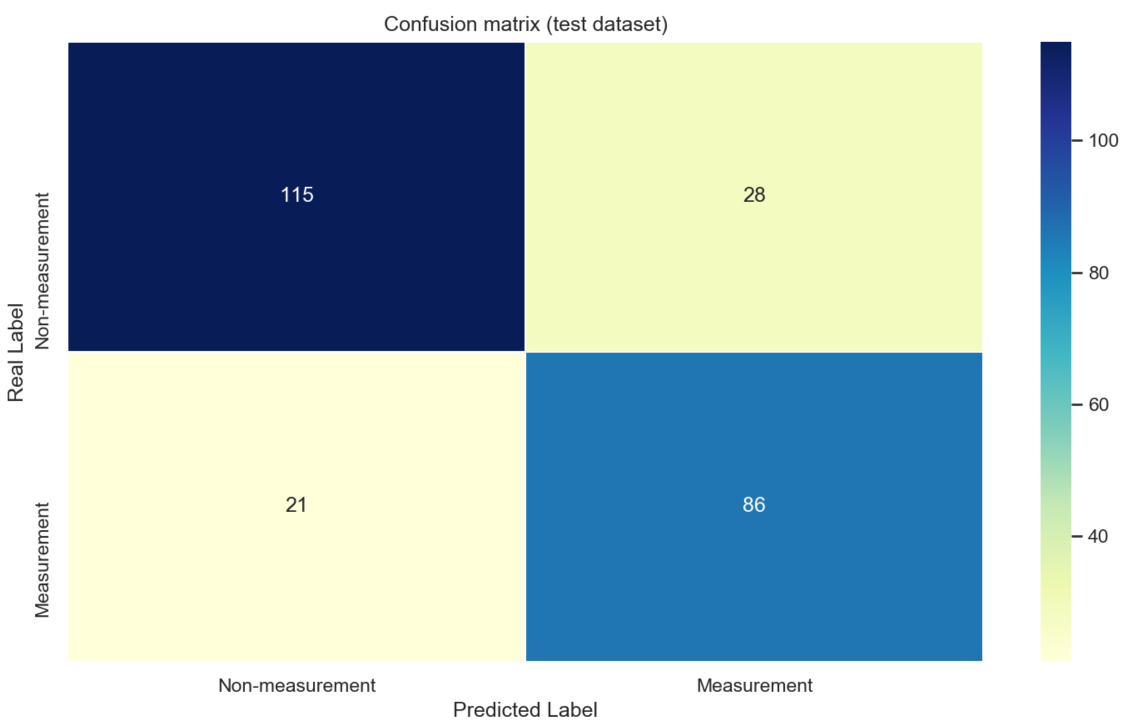

Figure 4 A confusion matrix of measurements and non measurements detected for interventions and crops

Table 3: Precision, Recall and F1-Measures for study design type			
	Precision	Recall	F1-measures
Non Measurement	0.85	0.8	0.82
Measurement	0.76	0.8	0.78
Averages (for precision, recall and F1-measures)	0.8	0.8	0.8

Table 3

Table Four. We present a consolidate table of all F1 —measures for ML tasks.

Table 4 Consolidated table of all F1 –measures for ML tasks	
Model	F1-measure
SVM-KNN-Gradient boosting model for interventions	0.85
Measurements BERT model	0.8
SciBERT fine-tuning model for study type	0.8

Table 4

Comparison of dataset size from other systematic reviews

Numbers of included studies that were reported in 20 systematic reviews in agriculture (author's analysis)			
Study title	**Total articles screened**	**Studies included**	**Included studies/total articles screened**
Gurwick, N. P., Moore, L. A., Kelly, C., & Elias, P. (2013). A systematic review of biochar research, with a focus on its stability in situ and its promise as a climate mitigation strategy. PloS one, 8(9).	**311**	**7**	**2%**
Crane, T. A., Delaney, A., Tamás, P. A., Chesterman, S., & Ericksen, P. (2017). A systematic review of local vulnerability to climate change in developing country agriculture. Wiley Interdisciplinary Reviews: Climate Change, 8(4), e464.	**168**	**38**	**23%**
Johnston, D., Stevano, S., Malapit, H. J., Hull, E., & Kadiyala, S. (2015). Agriculture, gendered time use, and nutritional outcomes: A systematic review.	**7000**	**23**	**0.3%**

Kamwamba-Mtethiwa, J., Weatherhead, K., & Knox, J. (2016). Assessing performance of small-scale pumped irrigation systems in sub-Saharan Africa: evidence from a systematic review. Irrigation and Drainage, 65(3), 308-318.	**1442**	**35**	**2%**
Ton, G., Vellema, W., Desiere, S., Weituschat, S., & D'Haese, M. (2018). Contract farming for improving smallholder incomes: What can we learn from effectiveness studies? World Development, 104, 46-64.	**8529**	**22**	**0.3%**
Ton, G., de Grip, K., Klerkx, L. W. A., Rau, M. L., Douma, M., Friis-Hansen, E., ... & Wongtschowski, M. (2013). Effectiveness of innovation grants to smallholder agricultural producers: an explorative systematic review. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.	**4322**	**20**	**0.5%**
Bodnár, F. (2011). Improving food security: a systematic review of the impact of interventions in agricultural production, value chains, market regulation, and land security.	**1269**	**46**	**4%**
DeFries, R. S., Fanzo, J., Mondal, P., Remans, R., & Wood, S. A. (2017). Is voluntary certification of tropical agricultural commodities achieving sustainability goals for small-scale producers? A review of the evidence. Environmental Research Letters, 12(3), 033001.	**2658**	**24**	**1%**
Cole, S., Bastian, G., Vyas, S., Wendel, C., & Stein, D. (2012). The effectiveness of index-based micro-insurance in helping smallholders manage weather-related risks. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, 59.	**5558**	**13**	**0.2%**
Girard, A. W., Self, J. L., McAuliffe, C., & Olude, O. (2012). The effects of household food production strategies on the health and nutrition outcomes of women and young children: a systematic review. Paediatric and Perinatal Epidemiology, 26, 205-222.	**3411**	**36**	**1%**
Sola, P., Cerutti, P. O., Zhou, W., Gautier, D., Iiyama, M., Schure, J., ... & Petrokofsky, G. (2017). The environmental, socioeconomic, and health impacts of woodfuel value chains in Sub-Saharan Africa: a systematic map. Environmental Evidence, 6(1), 4.	**4728**	**131**	**3%**
Persson, U. M. (2015). The impact of biofuel demand on agricultural commodity prices: a systematic review. Wiley Interdisciplinary Reviews: Energy and Environment, 4(5), 410-428.	**2068**	**121**	**6%**
Lawry, S., Samii, C., Hall, R., Leopold, A., Hornby, D., & Mtero, F. (2014). The impact of land property rights interventions on investment and agricultural productivity in developing countries: a systematic review. Campbell Systematic Reviews, 10(1), 1-104.	**27632**	**20**	**0.1%**
Loevinsohn, M., Sumberg, J., Diagne, A., & Whitfield, S. (2013). Under what circumstances and conditions does adoption of technology result in increased agricultural productivity? A Systematic Review.	**20299**	**5**	**0.02%**
Thorn, J. P., Friedman, R., Benz, D., Willis, K. J., & Petrokofsky, G. (2016). What evidence exists for the effectiveness of on-farm conservation land management strategies for preserving ecosystem services in developing countries? A systematic map. Environmental Evidence, 5(1), 13.	**21147**	**746**	**4%**
McCorriston, S., Hemming, D., Lamontagne-Godwin, J. D., Osborn, J., Parr, M. J., & Roberts, P. D. (2013). What is the evidence of the impact of agricultural trade liberalisation on food security in developing countries?. EPPI-Centre, Social Service Research Unit, Institute of Education, University of London.	**1176**	**34**	**3%**
Duvendack, M., Palmer-Jones, R., Copestake, J. G., Hooper, L., Loke, Y., & Rao, N. (2011). What is the evidence of the impact of microfinance on the well-being of poor people?.	**2643**	**74**	**3%**
MEDIAN PERCENTAGE of included studies across all reviews			**2%**