AUTO-TAGGING GLOBAL GIVING PROJECTS

BY: MELISSA BARONA

SPRINGBOARD CAPSTONE PROJECT 3

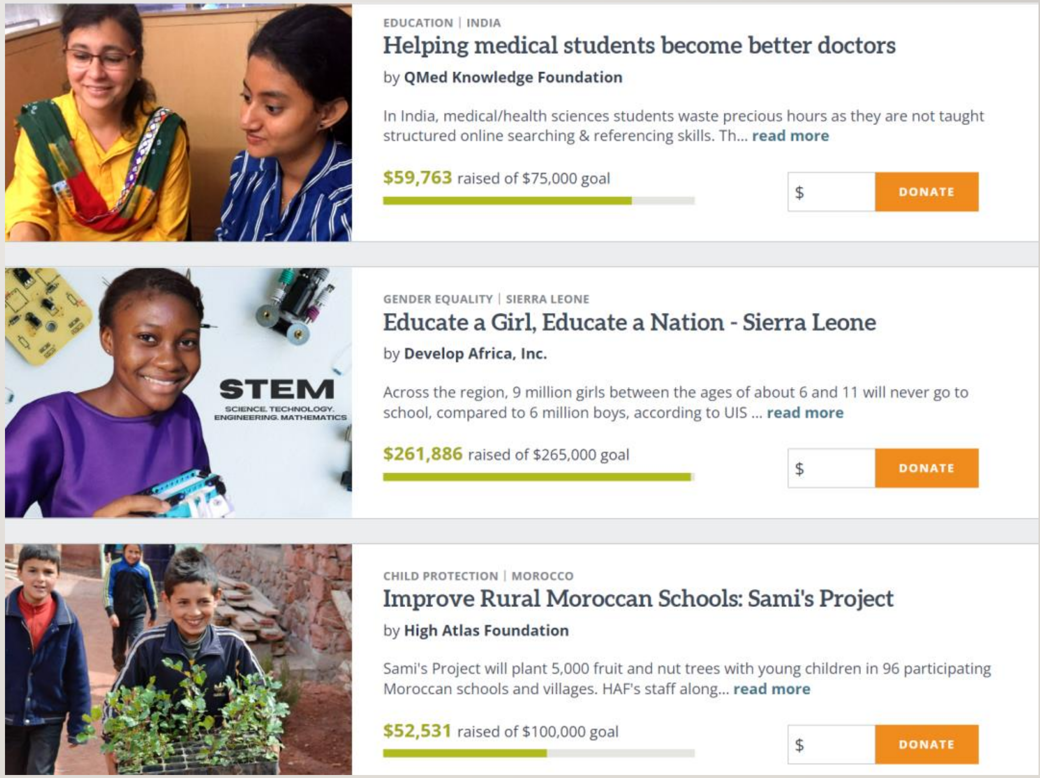
PROBLEM IDENTIFICATION

Problem:

It can be tedious, expensive, and time-consuming to manually assign each project to a theme

Solution:

Create an auto-tagging system to classify projects using machine learning



EDUCATION | INDIA

Helping medical students become better doctors

by **QMed Knowledge Foundation**

In India, medical/health sciences students waste precious hours as they are not taught structured online searching & referencing skills. Th... **read more**

$59,763 raised of $75,000 goal

$ DONATE

GENDER EQUALITY | SIERRA LEONE

Educate a Girl, Educate a Nation - Sierra Leone

by **Develop Africa, Inc.**

Across the region, 9 million girls between the ages of about 6 and 11 will never go to school, compared to 6 million boys, according to UIS ... **read more**

$261,886 raised of $265,000 goal

$ DONATE

CHILD PROTECTION | MOROCCO

Improve Rural Moroccan Schools: Sami's Project

by **High Atlas Foundation**

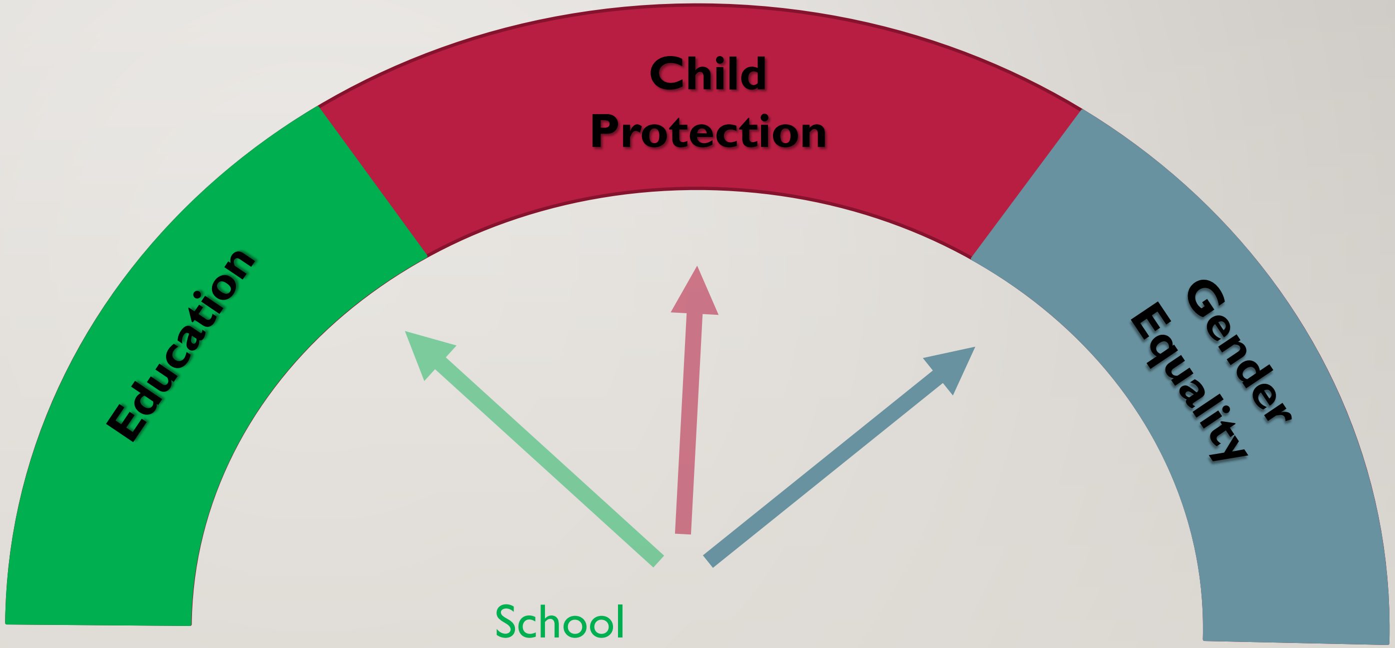
Sami's Project will plant 5,000 fruit and nut trees with young children in 96 participating Moroccan schools and villages. HAF's staff along... **read more**

$52,531 raised of $100,000 goal

$ DONATE

PRODUCT VISION

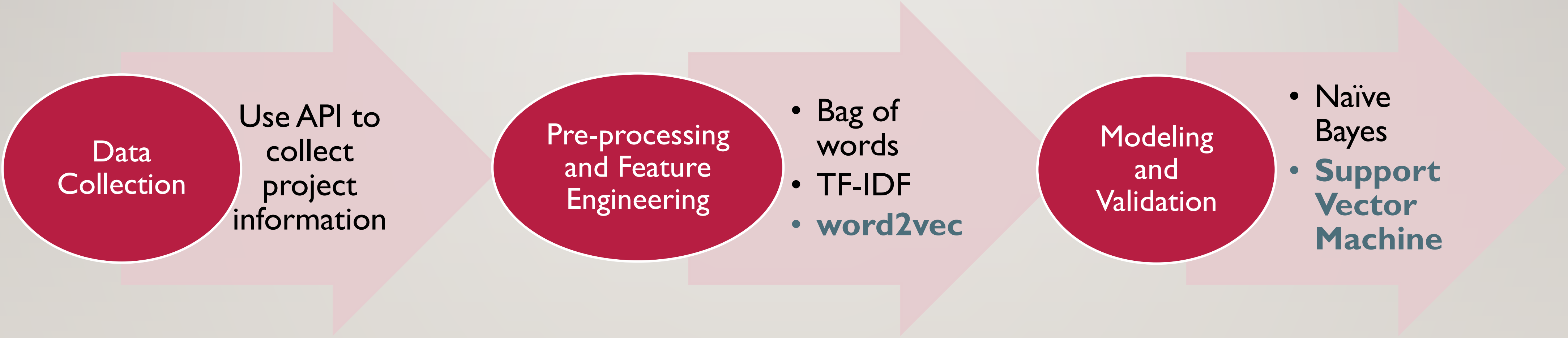
- An auto-tagger that properly tags a project based on the context of its description
- An appropriate classification of a project to a theme is essential for attracting the ideal donor that will continue to support the mission of the project.



School
Protect children in schools
girls in schools

OUR BEST CLASSIFIER:

A WORD2VEC TRANSFORMATION AND SUPPORT VECTOR MACHINE MODEL



Data Collection

Use API to collect project information

Pre-processing and Feature Engineering

- Bag of words
- TF-IDF
- **word2vec**

Modeling and Validation

- Naïve Bayes
- **Support Vector Machine**

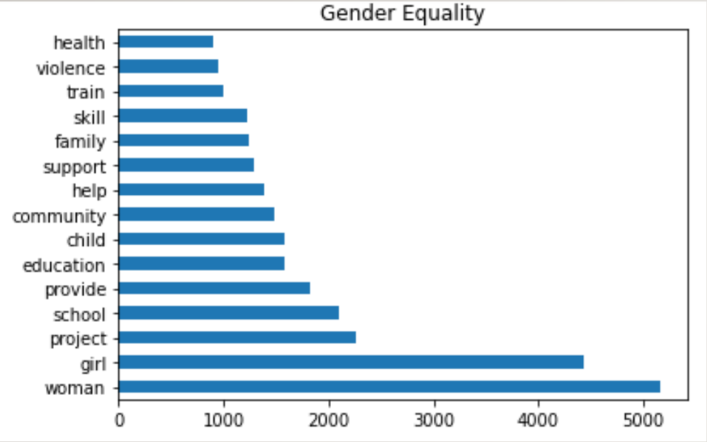
OUR BEST CLASSIFIER:

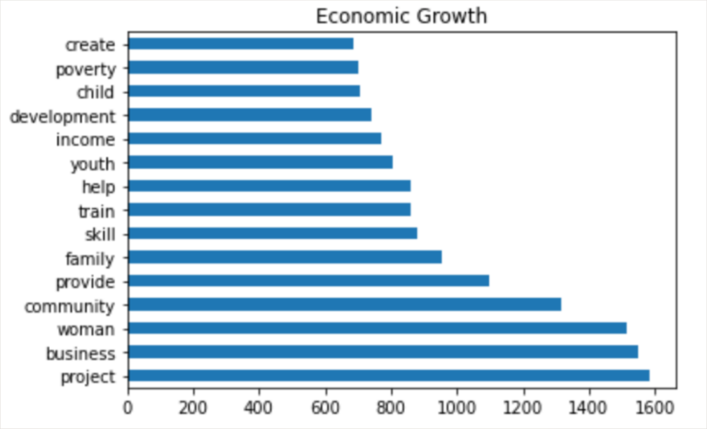
A WORD2VEC TRANSFORMATION AND SUPPORT VECTOR MACHINE MODEL

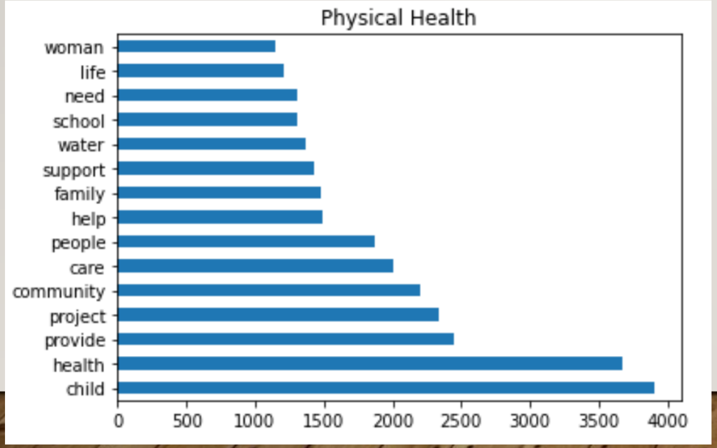
Model	Overall Accuracy	
	Training Set	Testing Set
TF-IDF/Naïve Bayes	0.94	0.92
TF-IDF/Linear SVM	0.94	0.84
Word2vec/Naïve Bayes	0.70	0.71
Word2vec/RDF SVM	**0.99**	**0.84**

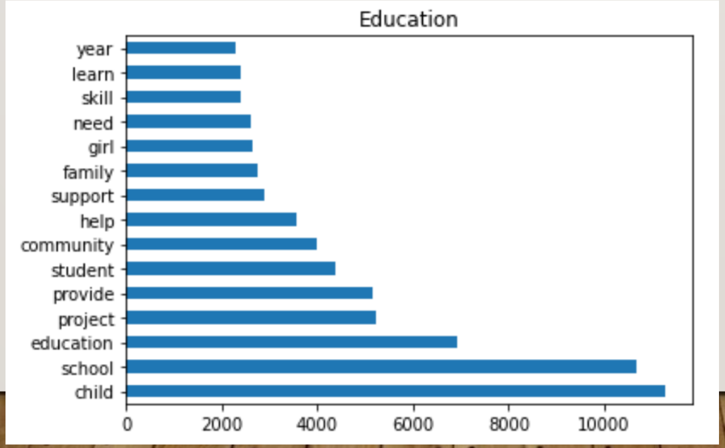
BUILDING THE AUTO-TAGGING CLASSIFIER:

WHAT WORDS WILL MATTER THE MOST WHEN CLASSIFYING?







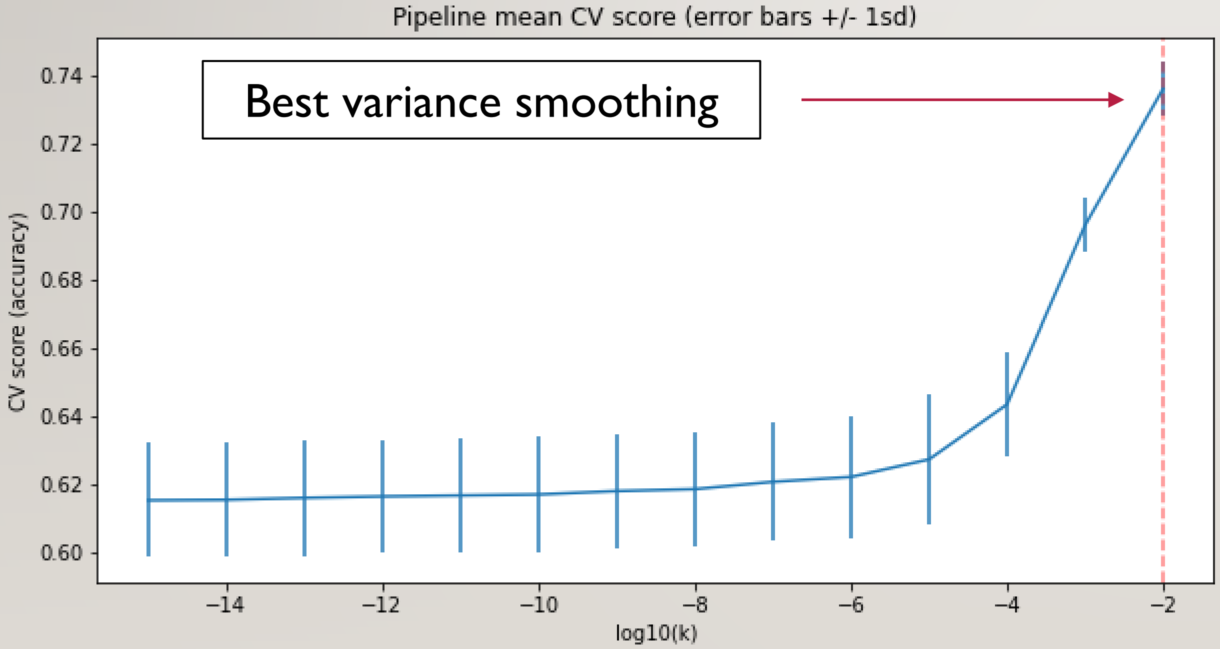


- ‘Project’, ‘provide’, ‘help’, ‘child’, ‘community’, and ‘family’ appear in all categories
- ‘School’, ‘education’, ‘woman’, and ‘girl’ appear often in both ‘Gender Equality’ and ‘Education’ texts which can lead to some texts being misclassified under these themes
- ‘Water’ and ‘health’ are most frequent in ‘Physical Health’ and ‘development’
- ‘income’ are most frequent in ‘Economic Growth’ texts

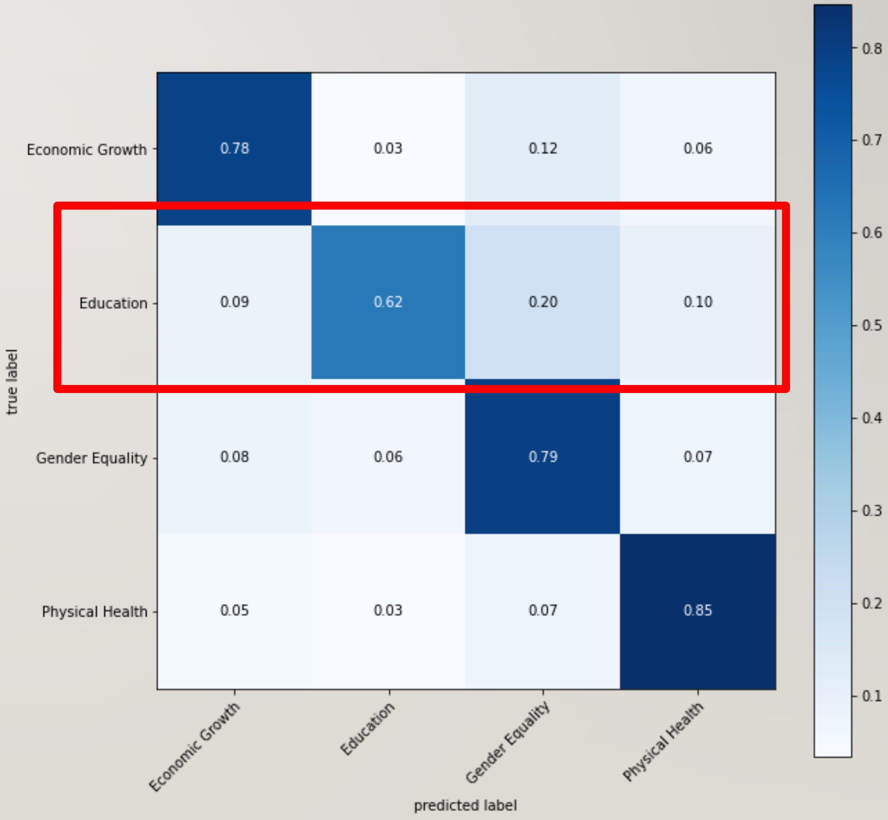
BUILDING THE AUTO-TAGGING CLASSIFIER:

THE TF-IDF/NAÏVE BAYES MODEL MISCLASSIFIES "EDUCATION" PROJECTS THE MOST

Hyper-parameter tuning


Pipeline mean CV score (error bars +/- 1sd)

Best variance smoothing

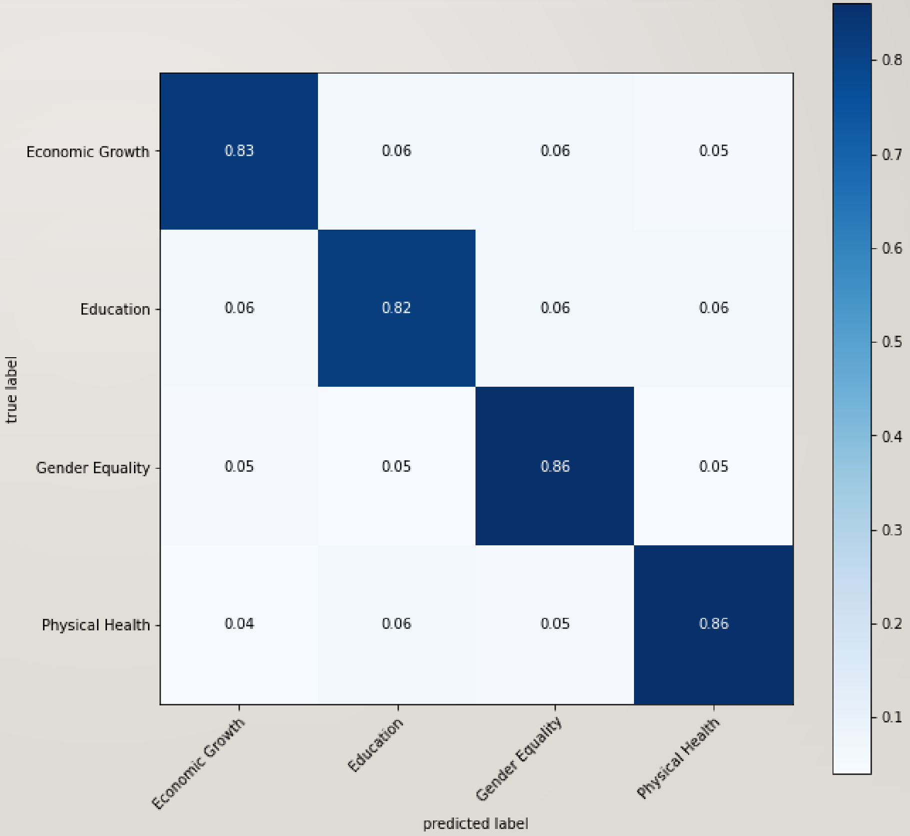


true label \ predicted label	Economic Growth	Education	Gender Equality	Physical Health
Economic Growth	0.78	0.03	0.12	0.06
Education	0.09	0.62	0.20	0.10
Gender Equality	0.08	0.06	0.79	0.07
Physical Health	0.05	0.03	0.07	0.85

BUILDING THE AUTO-TAGGING CLASSIFIER:

THE TF-IDF/SVM MODEL IS MORE ACCURATE, BUT MUCH SLOWER TO TRAIN

- TF-IDF/Linear SVM model correctly classifies 20% more “Education” projects than the TF-IDF/Naïve Bayes model

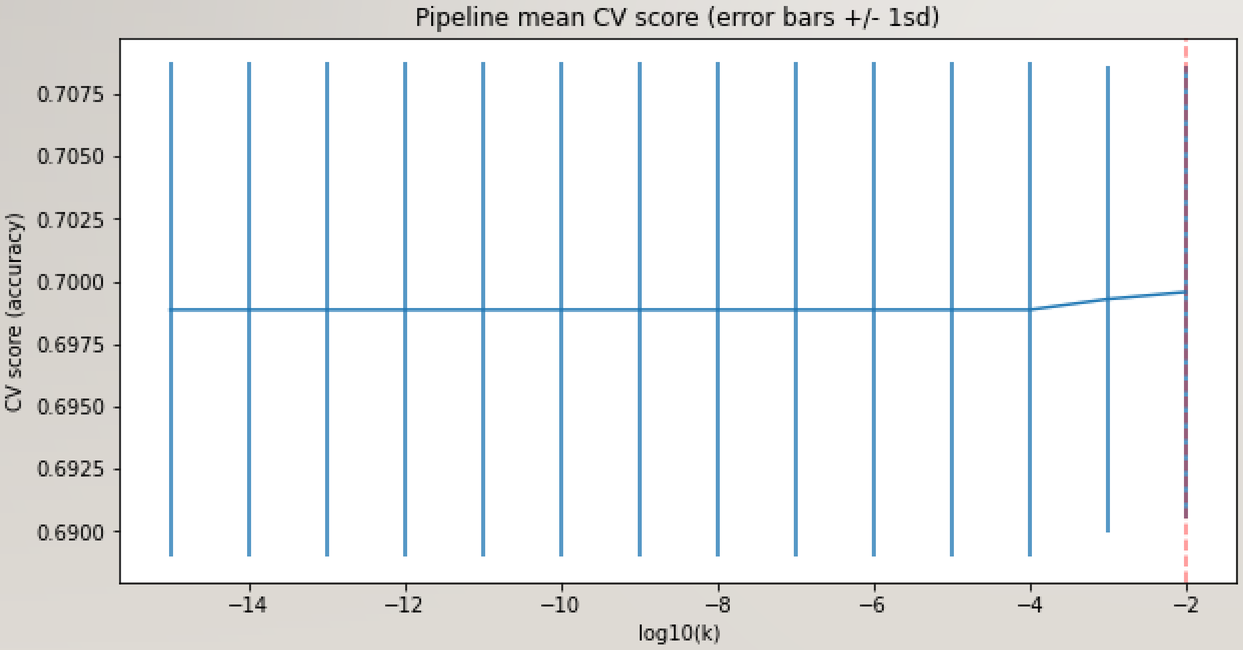


true label \ predicted label	Economic Growth	Education	Gender Equality	Physical Health
Economic Growth	0.83	0.06	0.06	0.05
Education	0.06	0.82	0.06	0.06
Gender Equality	0.05	0.05	0.86	0.05
Physical Health	0.04	0.06	0.05	0.86

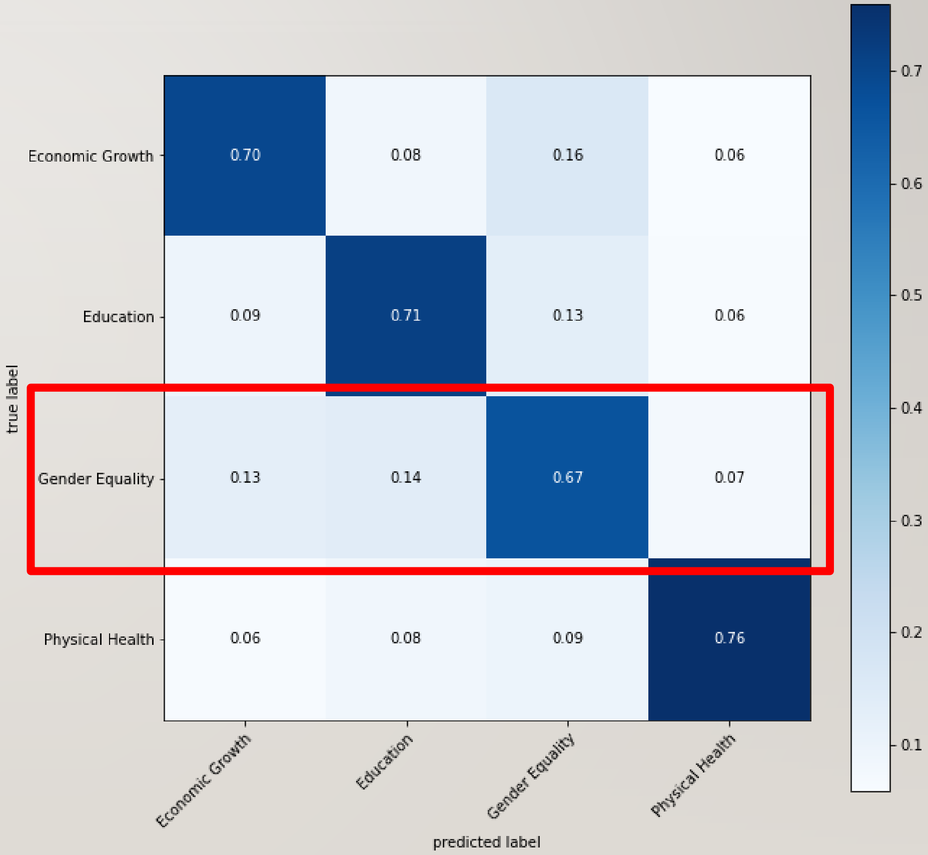
BUILDING THE AUTO-TAGGING CLASSIFIER:

THE WORD2VEC/NAÏVE BAYES MODEL MISCLASSIFIES “GENDER EQUALITY” PROJECTS THE MOST

Tuning the variance smoothing parameter has no effect on the performance of the model



Pipeline mean CV score (error bars +/- 1sd)

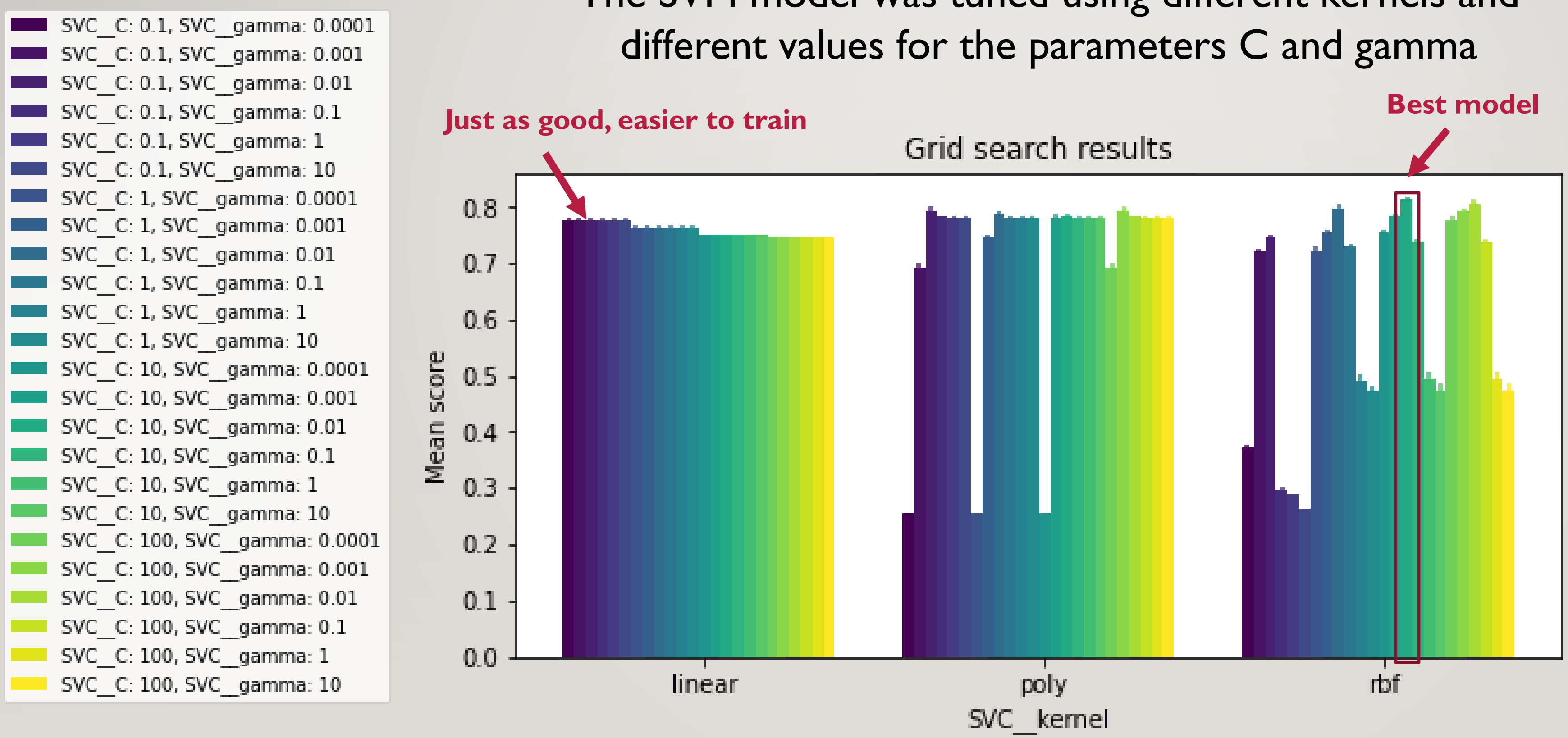


	Economic Growth	Education	Gender Equality	Physical Health
Economic Growth	0.70	0.08	0.16	0.06
Education	0.09	0.71	0.13	0.06
Gender Equality	0.13	0.14	0.67	0.07
Physical Health	0.06	0.08	0.09	0.76

THE MOST OPTIMAL CLASSIFIER:

A WORD2VEC/SVM MODEL

The SVM model was tuned using different kernels and different values for the parameters C and gamma



Just as good, easier to train

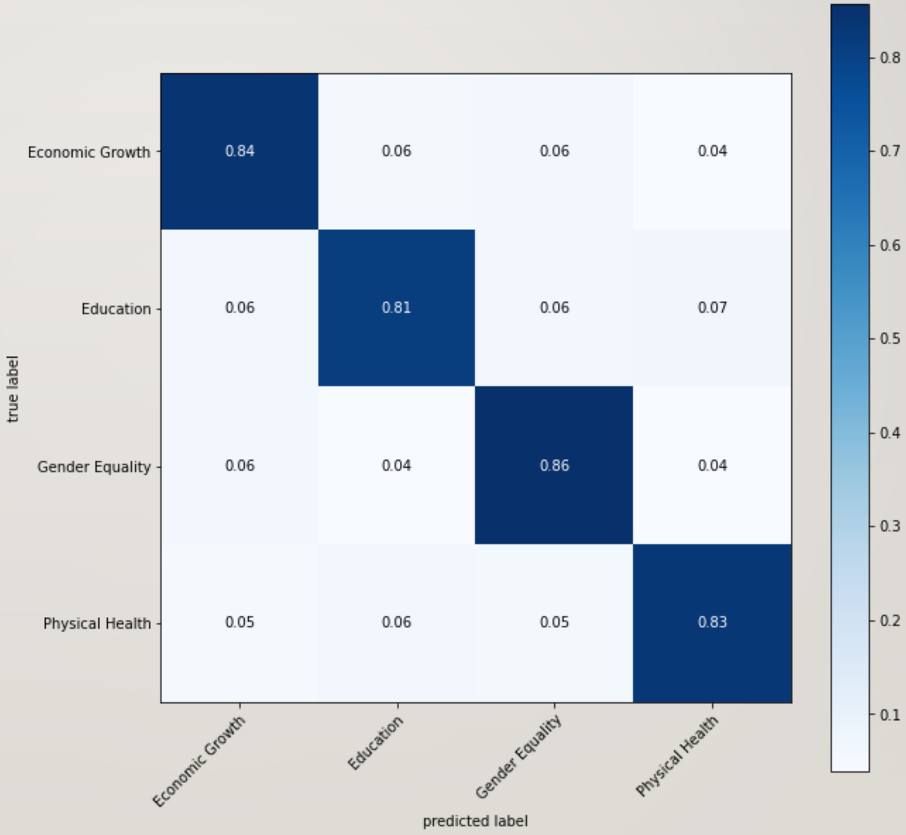
Best model

- The best performing kernel is RBF with parameter values C = 10 and gamma = 0.01
- A linear SVM model with C = 0.1 performs well with an accuracy of nearly 0.80

THE MOST OPTIMAL CLASSIFIER:

A WORD2VEC/SVM MODEL

- Word2vec/RBF SVM model correctly classifies 19% more “Gender Equality” projects than the word2vec/Naïve Bayes model



true label \ predicted label	Economic Growth	Education	Gender Equality	Physical Health
Economic Growth	0.84	0.06	0.06	0.04
Education	0.06	0.81	0.06	0.07
Gender Equality	0.06	0.04	0.86	0.04
Physical Health	0.05	0.06	0.05	0.83

FUTURE IMPROVEMENTS

- Apply the word2vec word-embedding method to the entire dataset of containing all 38,811 projects and fit a radial SVM model
- Depending on how accurate this model is, the challenges of this next step will continue to be distinguishing among projects with similar language and identifying a metric that can assess how correctly the model classifies these projects, but also gives merit to partially correct classification