Automating the Identification of Interdisciplinary Papers with Machine Learning

Bea Lu

Problem:

- Nuclear Astrophysics is a major interdisciplinary research field where its literature is diffused amongst 40 different journals.
- At MSU, JINA-CEE manages a virtual journal curating weekly issues of relevant publications.
- Filtering ~500 new articles per week takes a lot of time.

Goal:

- Train a text classification model to help identify relevant papers.

Method:

Gather and clean data from arXiv

↓

Use TF-IDF for text vectorization

↓

Train an SVM model for classification

↓

Make predictions with new data

TF-IDF

- Term Frequency–inverse Document Frequency
- Measures how relevant a word is to a document in a corpus

$$w_{t,d} = tf_{t,d} \times \log(\frac{N}{df_t})$$

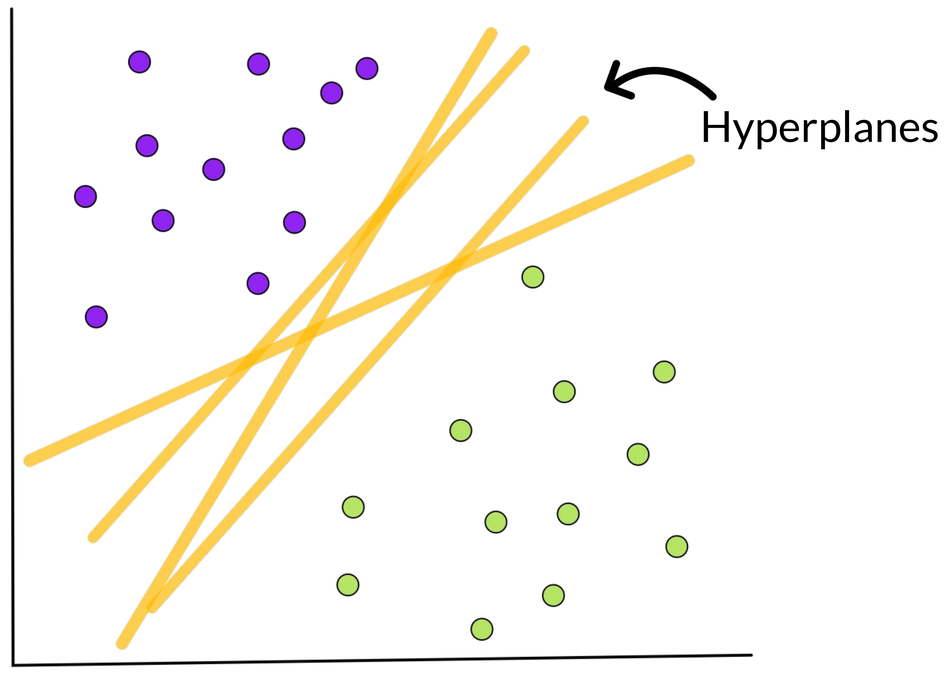
$tf_{t,d}$ = # of times term, t, shows up in a document, d

N = # of documents

df_t = # of documents where t shows up

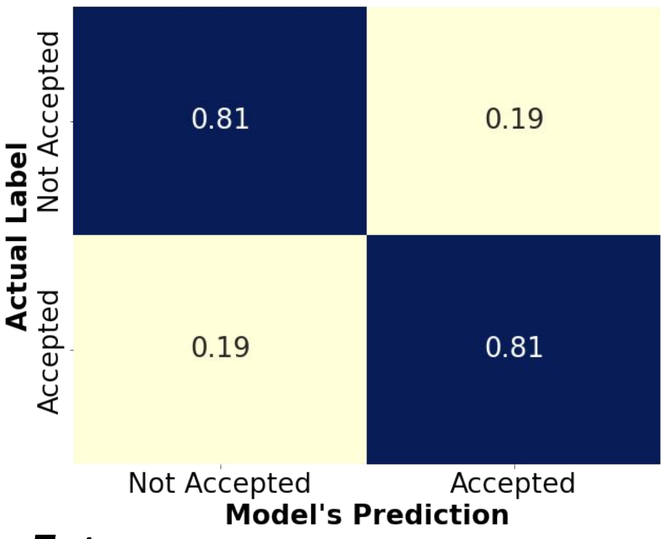
SVM Model

- Support Vector Machine
- Model utilizing hyperplanes for the separation of classes



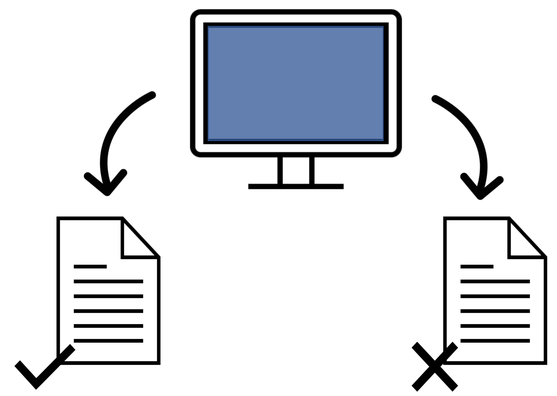
Results:

- Achieved 83% accuracy on 15,000 test papers.
- Only takes 18.5 ms per prediction.



Future:

- Revaluating training data for bias and errors
- Improving the accuracy of the current SVM model.
- Testing different models for comparisons.



Collaborators:
Vicente Amado Olivo
Wolfgang Kerzendorf

JINA-CEE