AUTOMATING GRANT CHARACTERISTICS USING NLP & MACHINE LEARNING

Helping to End Addiction Long-Term Initiative (HEAL)

Anthony Juehne — Program Officer and Data Scientist | HEAL Data Ecosystem

Erin Spaniol — Policy and Evaluation Lead | HEAL Initiative

CIVIC DIGITAL FELLOWSHIP

NIH

Noreen Mayat

Barnard College, Columbia University

Data Science

MOTIVATIONS

- HEAL seeks to improve both pain management and prevention tactics for opioid use disorder.

- Automating classification of HEAL awards for portfolio analysis will:
    - Significantly reduce the time burden of portfolio analysts within HEAL.

    - Highlight research themes, connect investigators studying aligned targets and interventions and determine promising areas for allocating research support.

CIVIC DIGITAL FELLOWSHIP

PROJECT GOALS

- Primary Outcome:
  - Classify if a study's primary outcome is Pain, OUD or Both.
  - Multi-Class

- Milestone
  - Classify if a study is/is not a milestone project.
  - Binary Classification

- Science Type
  - Classify a study's science type.
  - Multi-Class
  - Multi-Label

METHODS

Natural Language Processing

- Rule-based approaches.
- Uses key word ontologies to classify and label studies.
- 956 studies

Supervised Machine-Learning

- Science Type—broke each class into its own binary classification problem.
- Models used:
  - Random Forest
  - K-Nearest Neighbors
  - Logistic Regression
  - Support Vector Machine
- 956 studies

METHODS

- Pre-Processing:
    - Abstracts, specific aims, public health relevance cleaned for stop words.
- Filtration:
    - Only preserved sentences with keywords.
- Regular Expressions:
    - Iterates by row.
    - Search for regexes related to each category in filtered columns.
    - Add found terms to individual lists (Pain vs. OUD).
- Labeling:
    - Determine which list has most terms → assign label.

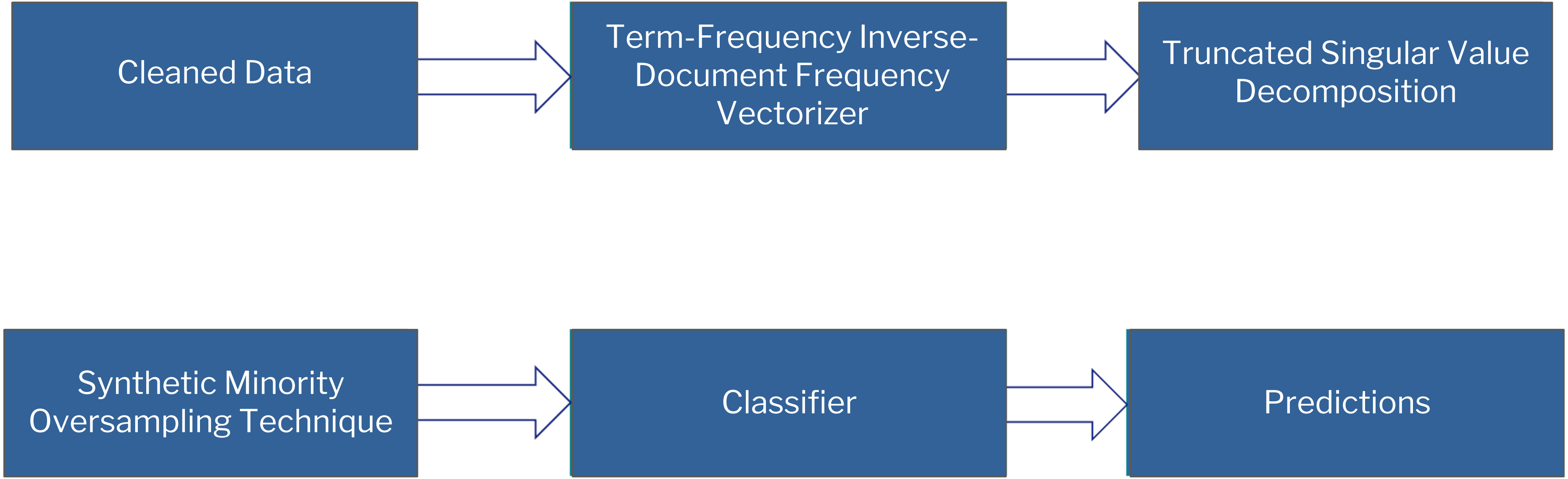
CIVIC DIGITAL FELLOWSHIP

METHODS

- Example Text: “although health social economic impacts **opioid addiction**...”
- oud_terms = [‘opioid addiction’]
- pain_terms = []
- both = []
- Study Outcome → OUD

CIVIC DIGITAL FELLOWSHIP

METHODS


Cleaned Data
Term-Frequency Inverse-Document Frequency Vectorizer
Truncated Singular Value Decomposition
Synthetic Minority Oversampling Technique
Classifier
Predictions

METHODS

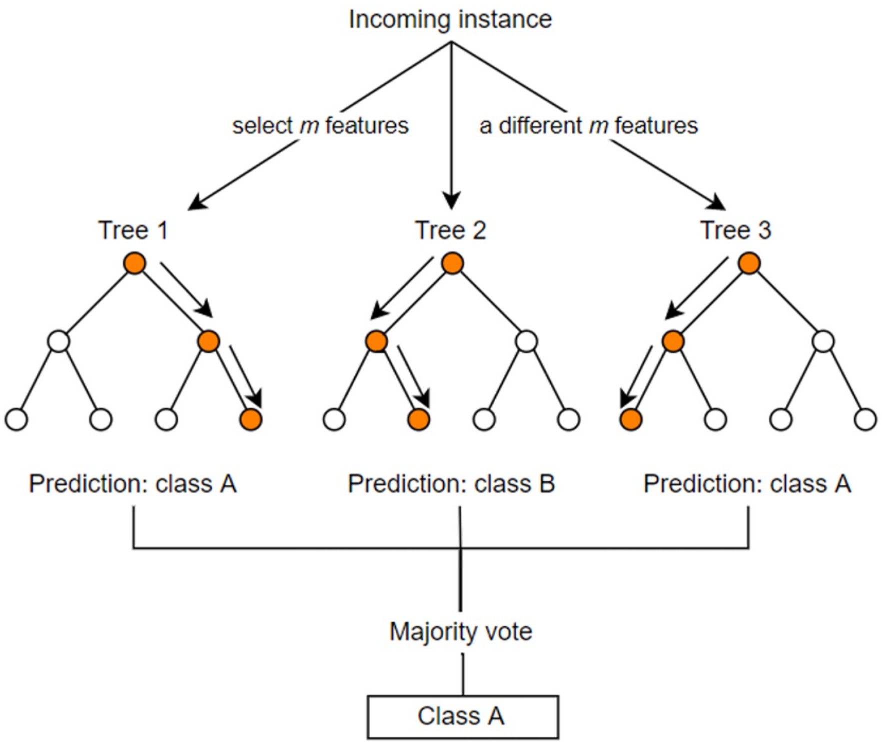
- Term Frequency-Inverse Document Frequency (TF-IDF) Matrix
- Normalized count of each word / Number of docs it appears in
- The higher the TF-IDF score the more important or relevant the term is

term	weight
pain	0.073881
opioid	0.054544
treatment	0.037505
oud	0.034639
use	0.03379
research	0.031719
care	0.031138
clinical	0.030721
health	0.02889
patients	0.026607
ctn	0.025051
chronic	0.023289
aim	0.022916
study	0.022679
phase	0.021094

CIVIC DIGITAL FELLOWSHIP

METHODS

- Random Forest:
  - Averages predictions of various decision trees.
  - Each root decision tree corresponds to a feature (word) in the study text → trickles down to a label.



CIVIC DIGITAL FELLOWSHIP

DATA

Fed to Classifier

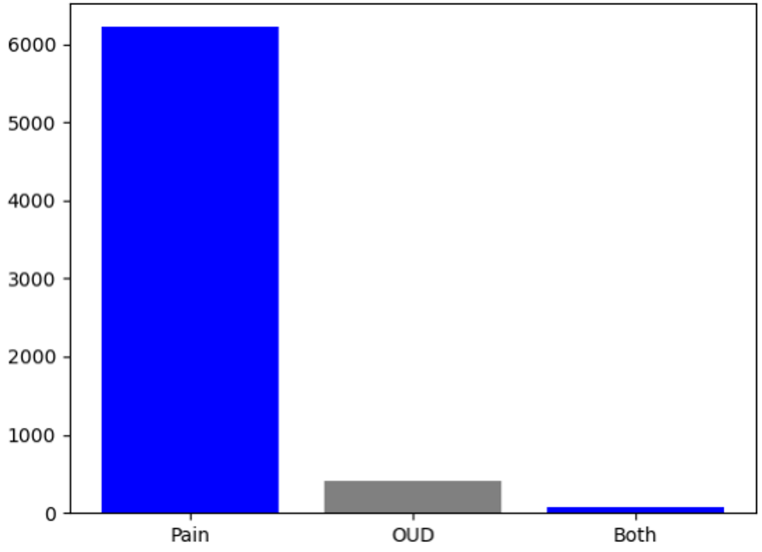
Designated Labels

	Appl ID	Combined Cleaned	HEAL Category- Primary Outcome
0	10459783	neonatal opioid withdrav	OUD
1	10133699	critical persistent gaps ev	Pain
2	10377726	number infants exposed	OUD
3	10378942	neonatal opioid withdrav	OUD
4	10378979	thomas jefferson univers	Both
5	10379584	neonatal opioid withdrav	OUD

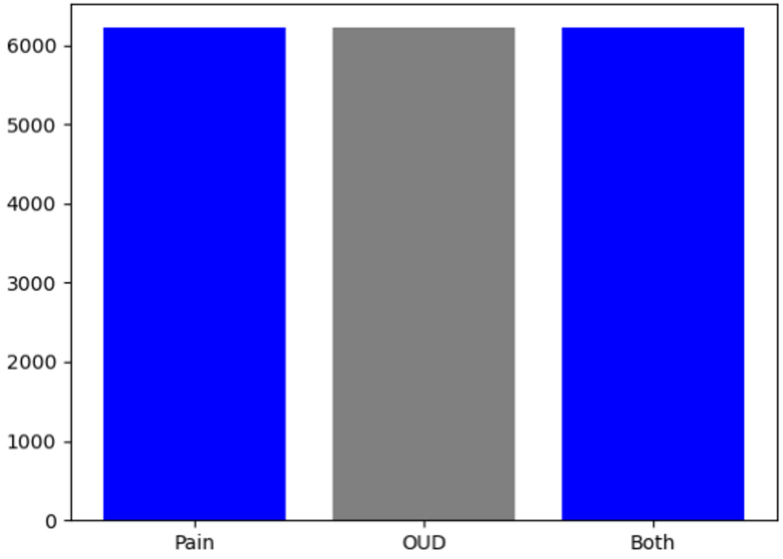
CIVIC DIGITAL FELLOWSHIP

DATA

Before SMOTE
Primary Outcome Data Distribution

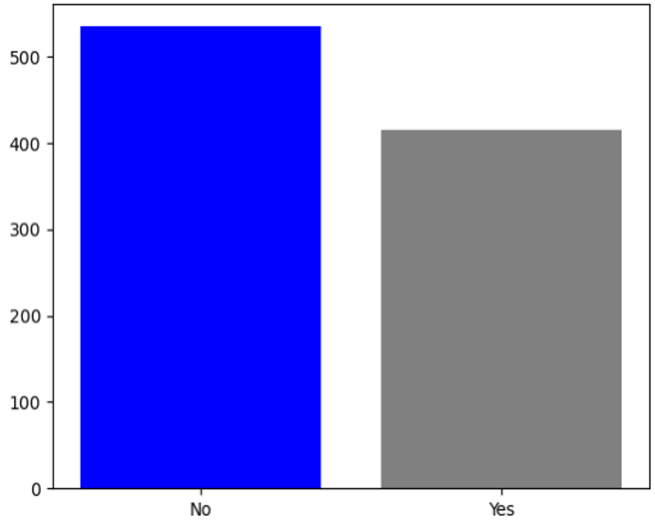


After SMOTE
Primary Outcome Data Distribution

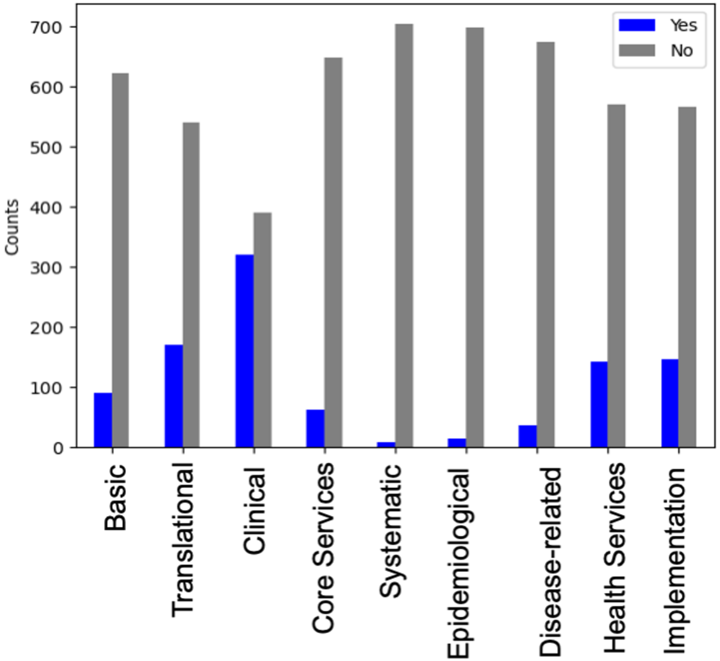


DATA

Milestone Distribution



Science Distribution



CIVIC DIGITAL FELLOWSHIP

RESULTS

- Primary Outcome:
  - Regex: 85%
  - Random Forest: 98%
- Milestone:
  - Regex: 76%
  - Random Forest: 84%

- Science Type:
  - KNN Basic: 92%
  - KNN Health Services Research: 85%
  - KNN Implementation Research: 80%

  - LR Disease-Related Basic: 88%
  - LR Clinical: 73%

  - RF Translational: 84%
  - RF Systematic Meta-analyses: 96%

  - SVM Core Services: 86%
  - SVM Epidemiological: 96%

CIVIC DIGITAL FELLOWSHIP

CONCLUSIONS & NEXT STEPS

- Expand OUD/Both datasets for primary outcome algorithm
- Hyperparameter tuning for ML algorithms
- NLP combined with ML approaches for final labeling
- Work combined in Jupyter notebook as well as Github for future building