FELLOW SUMMER 2022 PROJECTS

Dashboarding & Stem Cell Classification

Julia Craciun

CIVIC DIGITAL FELLOWSHIP | NIH

Jake Basson | Andrew Miklos | Jessie Wang
Nathan Moore | Jordan Jomsky

PROJECT COMPONENTS

3 main tasks

- Create an interactive portfolio overview dashboard of NIGMS grant funding practices in Tableau

- Assist in current dashboarding efforts

- Help Nate and Jordan build a classifier for stem cell grants

PORTFOLIO OVERVIEW DASHBOARD

Goal

- NIGMS 5-year strategic plan calls for monitoring and evaluating activities across the institute

- Very broad, exploratory purpose

- Allow NIGMS funding to be more equitable and visualize areas of improvement in the grant funding process

PORTFOLIO OVERVIEW DASHBOARD

Approach

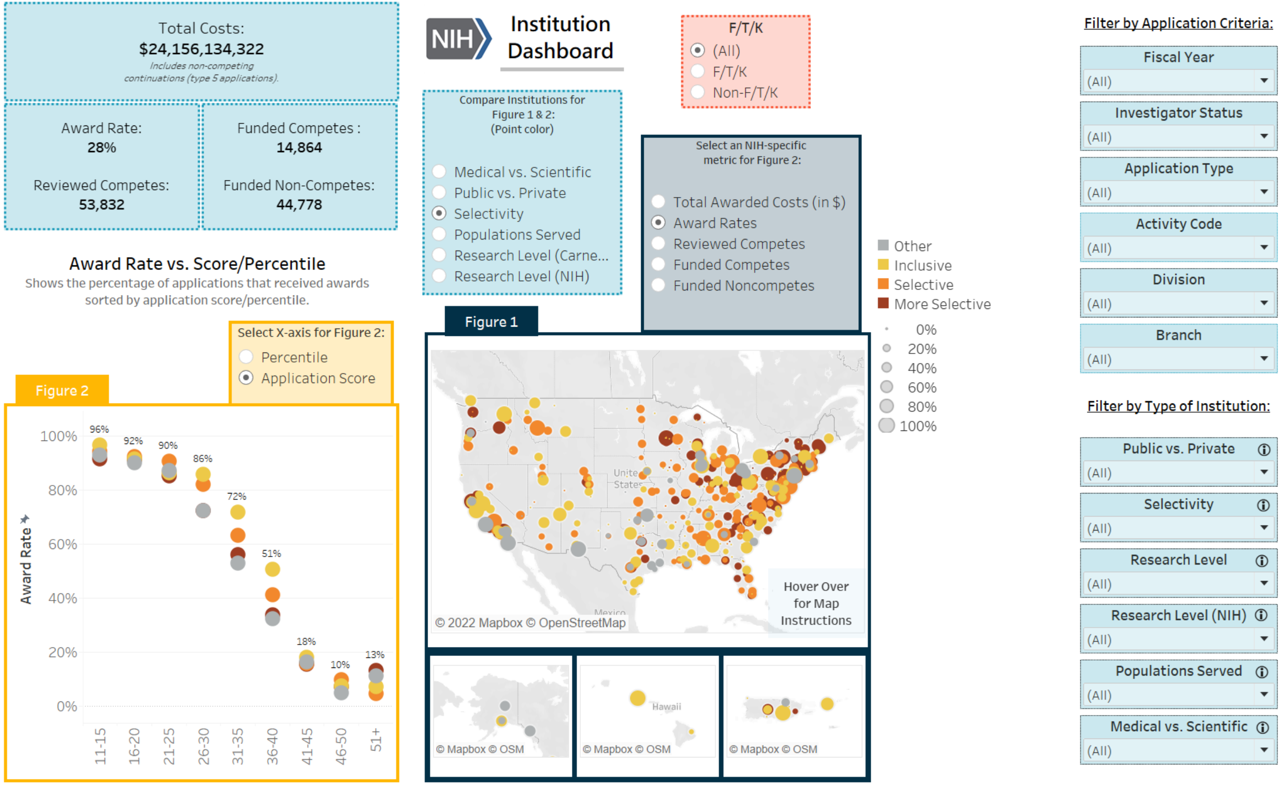
- Challenge: maintain simplicity and intuitiveness while visualizing variables of interest for a non-specific audience

- Consistent meetings and feedback regarding changes

- Noting places of confusion and inefficiency

- Determining variables of interest

RESULTS



NIH Institution Dashboard

Total Costs:
$24,156,134,322
Includes non-competing continuations (type 5 applications).

Award Rate: **28%**

Reviewed Competes: **53,832**

Funded Competes : **14,864**

Funded Non-Competes: **44,778**

Compare Institutions for Figure 1 & 2:
(Point color)

○ Medical vs. Scientific
○ Public vs. Private
● Selectivity
○ Populations Served
○ Research Level (Carne...
○ Research Level (NIH)

F/T/K
● (All)
○ F/T/K
○ Non-F/T/K

Select an NIH-specific metric for Figure 2:

○ Total Awarded Costs (in $)
● Award Rates
○ Reviewed Competes
○ Funded Competes
○ Funded Noncompetes

Award Rate vs. Score/Percentile

Shows the percentage of applications that received awards sorted by application score/percentile.

Select X-axis for Figure 2:
○ Percentile
● Application Score

■ Other
■ Inclusive
■ Selective
■ More Selective

● 0%
● 20%
● 40%
● 60%
● 80%
● 100%

Filter by Application Criteria:

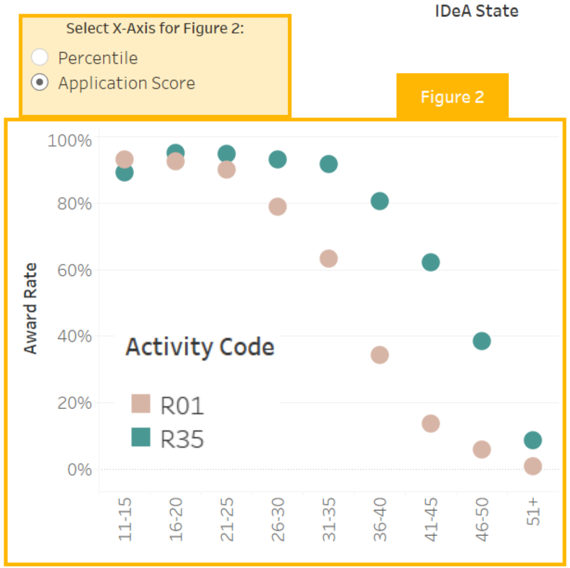
Fiscal Year
(All)
Investigator Status
(All)
Application Type
(All)
Activity Code
(All)
Division
(All)
Branch
(All)

Filter by Type of Institution:

Public vs. Private ⓘ
(All)
Selectivity ⓘ
(All)
Research Level ⓘ
(All)
Research Level (NIH) ⓘ
(All)
Populations Served ⓘ
(All)
Medical vs. Scientific ⓘ
(All)

CIVIC DIGITAL FELLOWSHIP

RESULTS



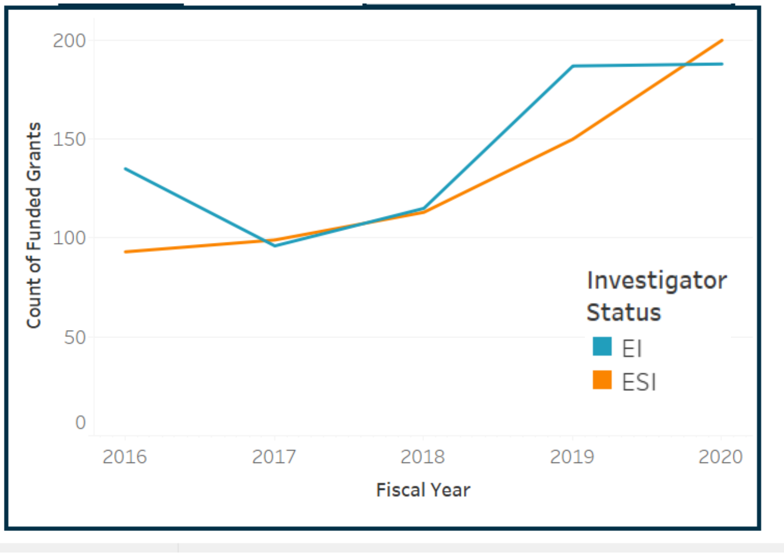
Select X-Axis for Figure 2:

○ Percentile

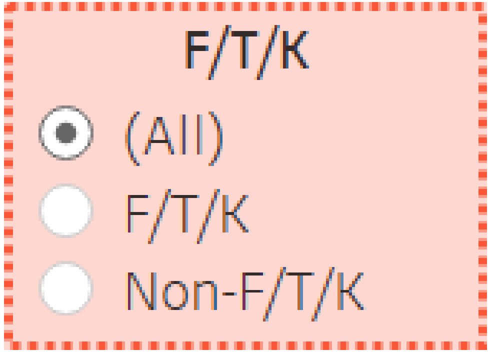
● Application Score

IDeA State

Figure 2



IMPROVEMENTS

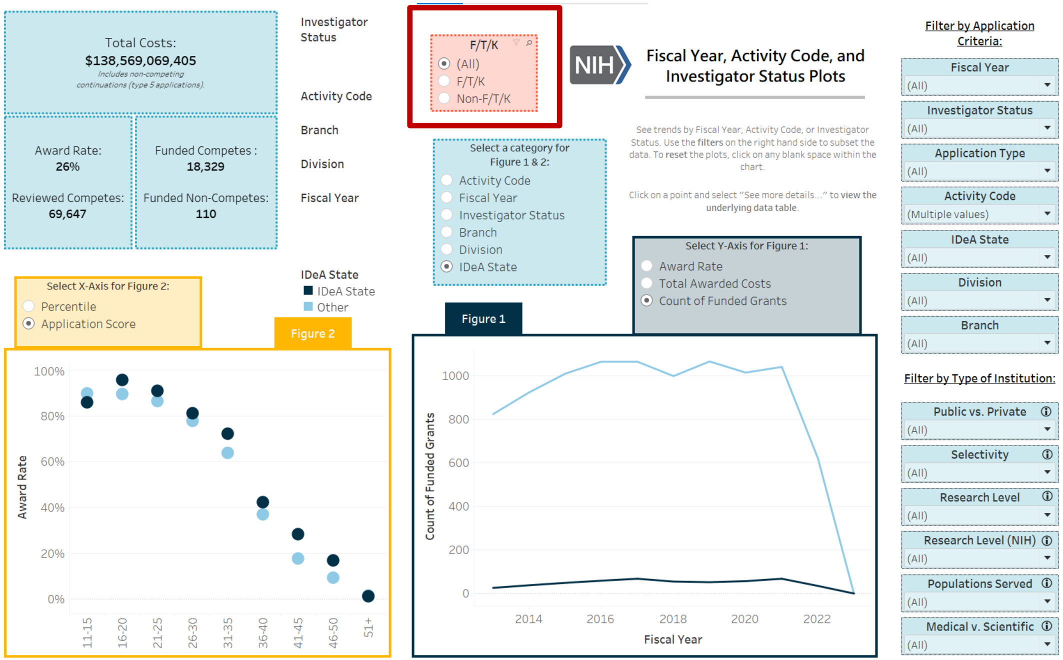


F/T/K

● (All)

○ F/T/K

○ Non-F/T/K



Total Costs:
$138,569,069,405
Includes non-competing continuations (type 5 applications).

Award Rate:
26%

Reviewed Competes:
69,647

Funded Competes :
18,329

Funded Non-Competes:
110

Investigator Status

Activity Code

Branch

Division

Fiscal Year

IDeA State

■ IDeA State

■ Other

F/T/K

● (All)

○ F/T/K

○ Non-F/T/K

Fiscal Year, Activity Code, and Investigator Status Plots

See trends by Fiscal Year, Activity Code, or Investigator Status. Use the **filters** on the right hand side to subset the data. To **reset** the plots, click on any blank space within the chart.

Click on a point and select "See more details..." to **view the underlying data table.**

Select a category for Figure 1 & 2:

○ Activity Code

○ Fiscal Year

○ Investigator Status

○ Branch

○ Division

● IDeA State

Select Y-Axis for Figure 1:

○ Award Rate

○ Total Awarded Costs

● Count of Funded Grants

Select X-Axis for Figure 2:

○ Percentile

● Application Score

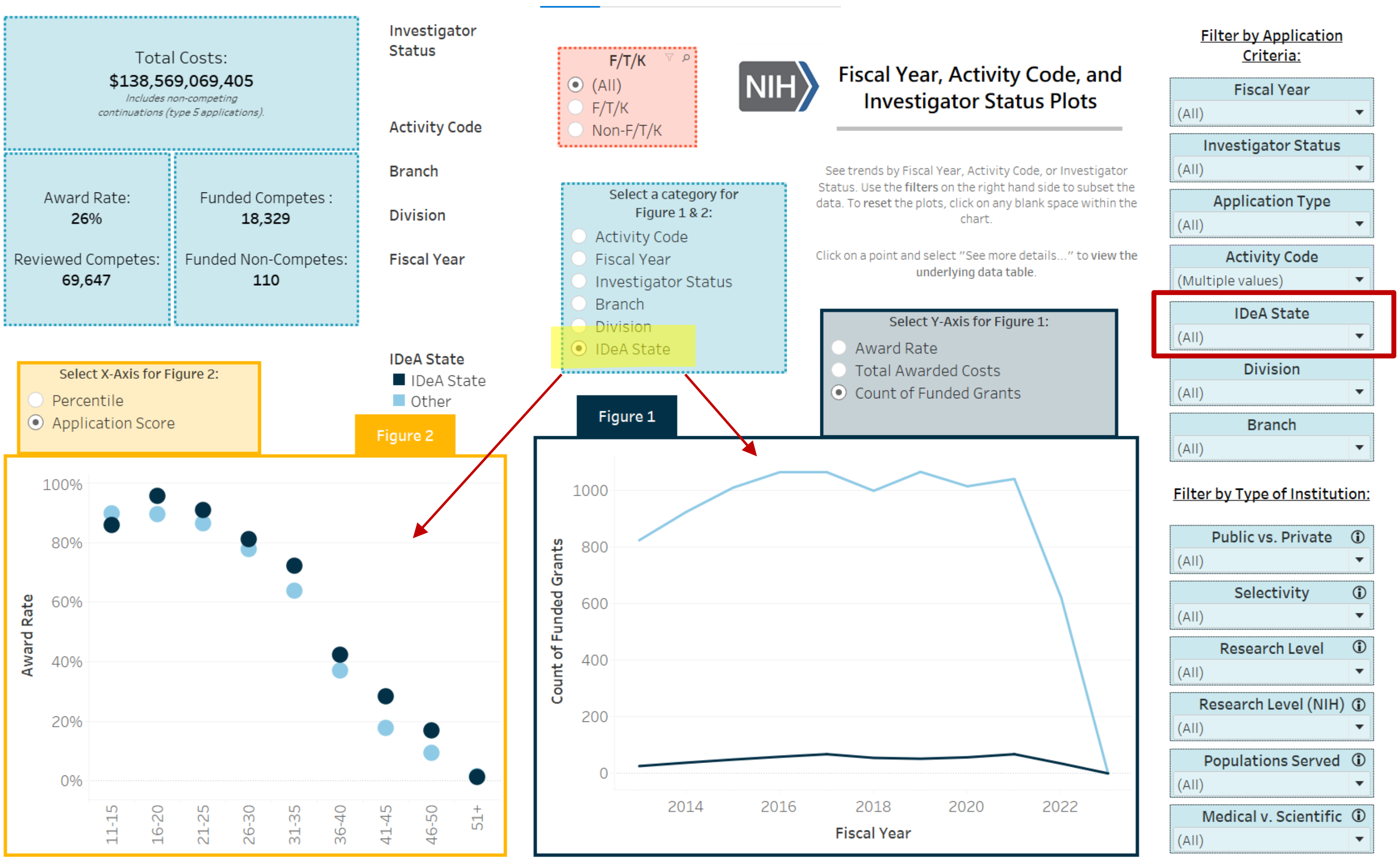
Filter by Application Criteria:

Fiscal Year	(All)
Investigator Status	(All)
Application Type	(All)
Activity Code	(Multiple values)
IDeA State	(All)
Division	(All)
Branch	(All)

Filter by Type of Institution:

Public vs. Private	(All)
Selectivity	(All)
Research Level	(All)
Research Level (NIH)	(All)
Populations Served	(All)
Medical v. Scientific	(All)

CIVIC DIGITAL FELLOWSHIP



Fiscal Year, Activity Code, and Investigator Status Plots

Total Costs:
$138,569,069,405
Includes non-competing continuations (type 5 applications).

Award Rate:
26%

Funded Competes :
18,329

Reviewed Competes:
69,647

Funded Non-Competes:
110

Investigator Status

Activity Code

Branch

Division

Fiscal Year

F/T/K
● (All)
○ F/T/K
○ Non-F/T/K

Select a category for Figure 1 & 2:
○ Activity Code
○ Fiscal Year
○ Investigator Status
○ Branch
○ Division
● IDeA State

See trends by Fiscal Year, Activity Code, or Investigator Status. Use the **filters** on the right hand side to subset the data. To **reset** the plots, click on any blank space within the chart.

Click on a point and select "See more details..." to **view the underlying data table**.

Select X-Axis for Figure 2:
○ Percentile
● Application Score

IDeA State
■ IDeA State
■ Other

Select Y-Axis for Figure 1:
○ Award Rate
○ Total Awarded Costs
● Count of Funded Grants

Figure 2

Figure 1

Filter by Application Criteria:

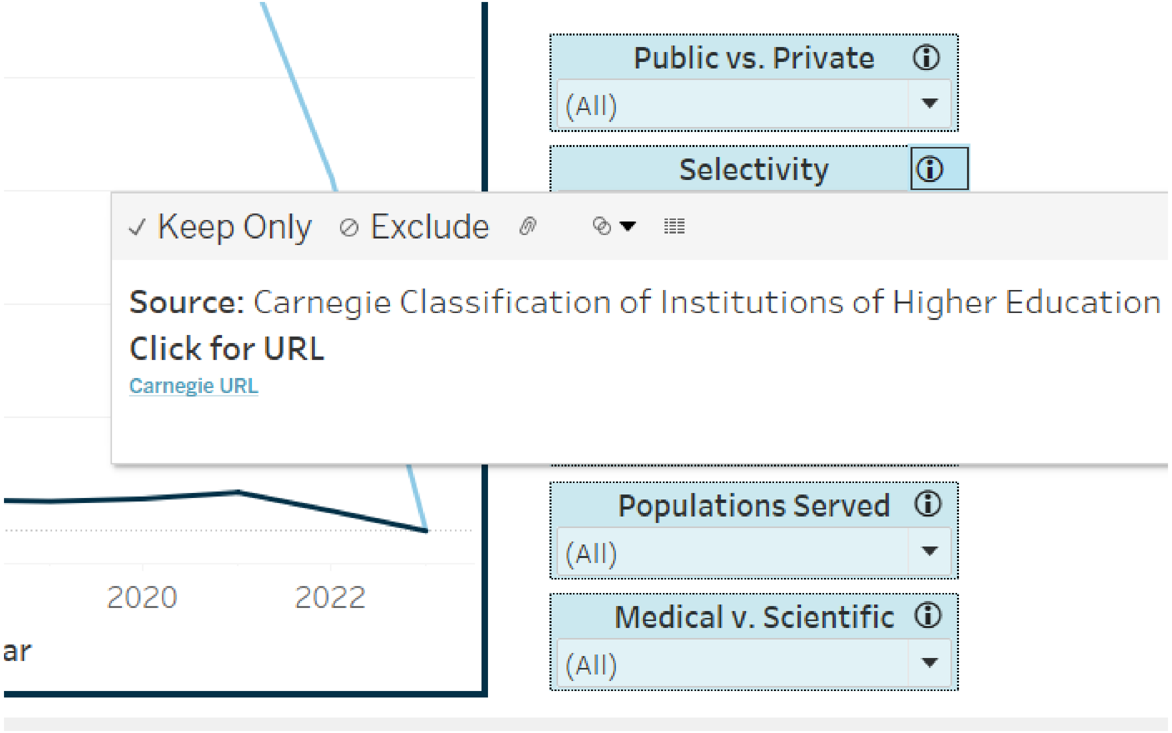
Fiscal Year	(All)
Investigator Status	(All)
Application Type	(All)
Activity Code	(Multiple values)
IDeA State	(All)
Division	(All)
Branch	(All)

Filter by Type of Institution:

Public vs. Private	(All)
Selectivity	(All)
Research Level	(All)
Research Level (NIH)	(All)
Populations Served	(All)
Medical v. Scientific	(All)

CIVIC
DIGITAL
FELLOWSHIP

IMPROVEMENTS



Source: Carnegie Classification of Institutions of Higher Education

Click for URL

Carnegie URL

Redacted Slide: Other Dashboarding Efforts

Redacted Slide: Other Dashboarding Efforts

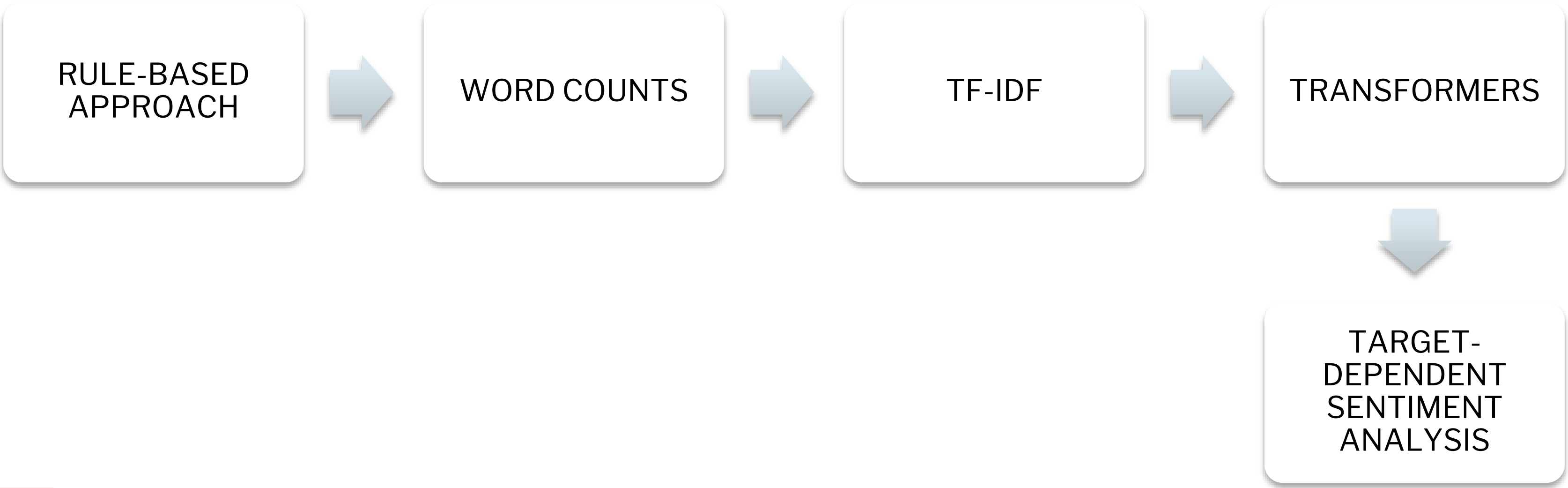
STEM CELL CLASSIFIER

Goal

- Assist Nate and Jordan in building a model that could classify the stem cell lines used within grants
- Some grants had been misclassified in the past
- Provide a tool that improves efficiency of stem cell classification
- Desire to assist POs in classifying these grants as human vs. non-human:
  - ○ Embryonic
  - ○ Non-embryonic
  - ○ Induced pluripotent
  - ○ Umbilical / placenta

STEM CELL CLASSIFIER

“The less data science you use, the better”



RULE-BASED APPROACH

WORD COUNTS

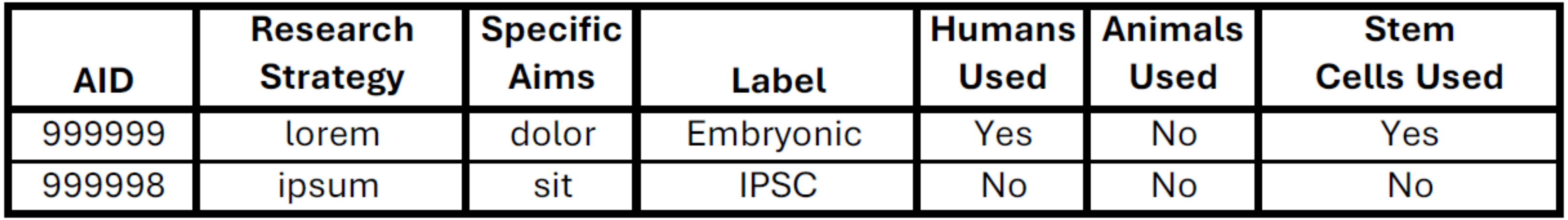
TF-IDF

TRANSFORMERS

TARGET-DEPENDENT SENTIMENT ANALYSIS

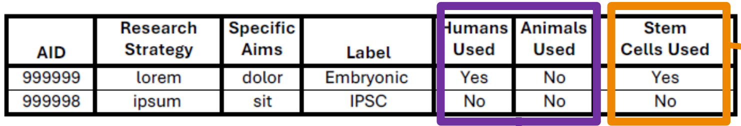
1. RULE-BASED APPROACH

Data

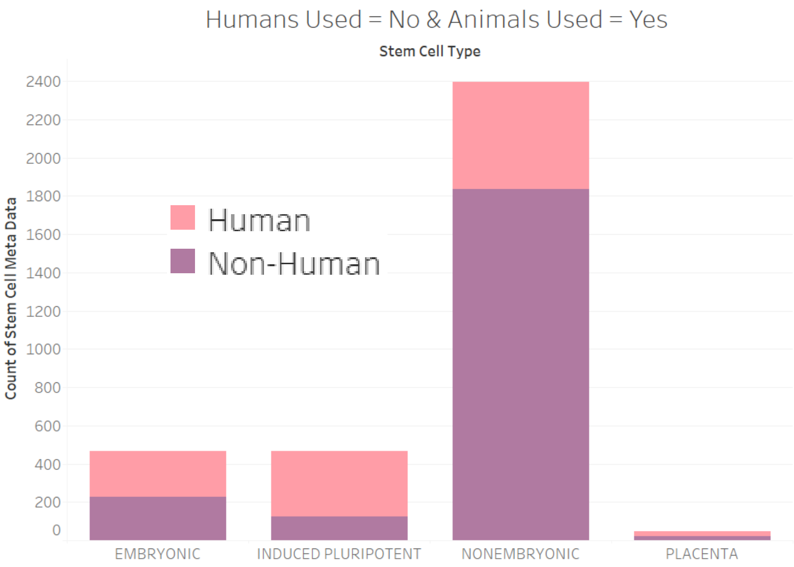


AID	Research Strategy	Specific Aims	Label	Humans Used	Animals Used	Stem Cells Used
999999	lorem	dolor	Embryonic	Yes	No	Yes
999998	ipsum	sit	IPSC	No	No	No

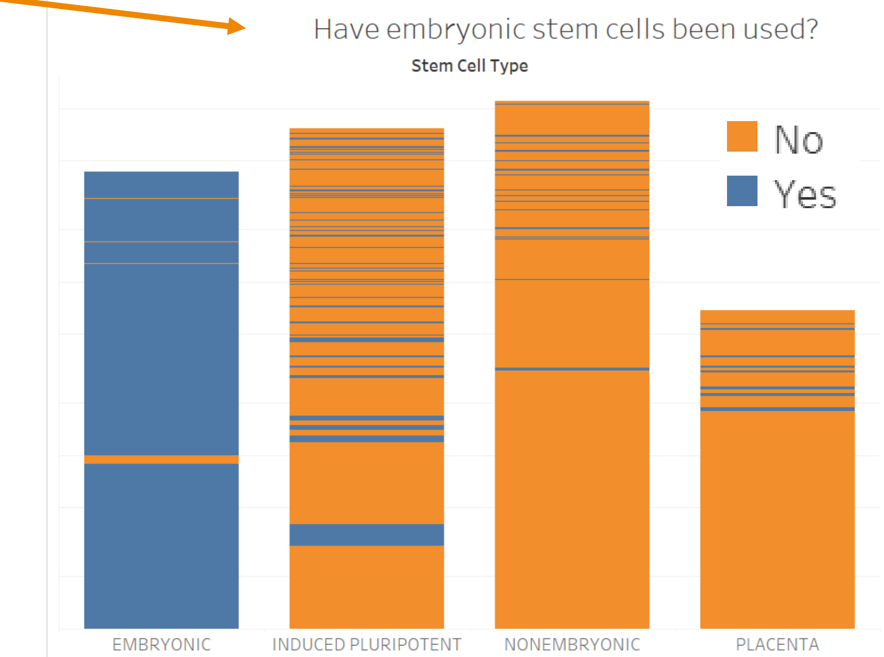
1. RULE-BASED APPROACH



AID	Research Strategy	Specific Aims	Label	Humans Used	Animals Used	Stem Cells Used
999999	lorem	dolor	Embryonic	Yes	No	Yes
999998	ipsum	sit	IPSC	No	No	No

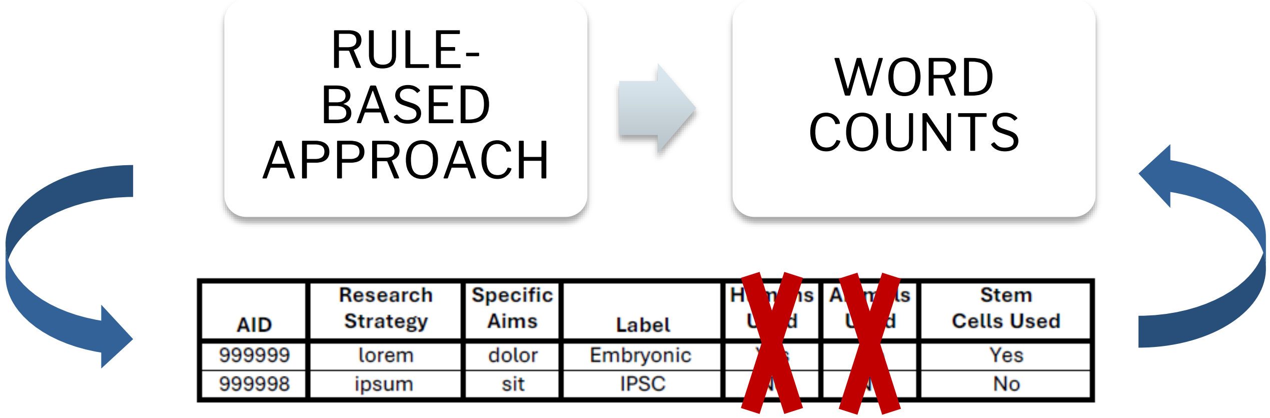


Humans Used = No & Animals Used = Yes



Have embryonic stem cells been used?

BACK TO THE DRAWING BOARD

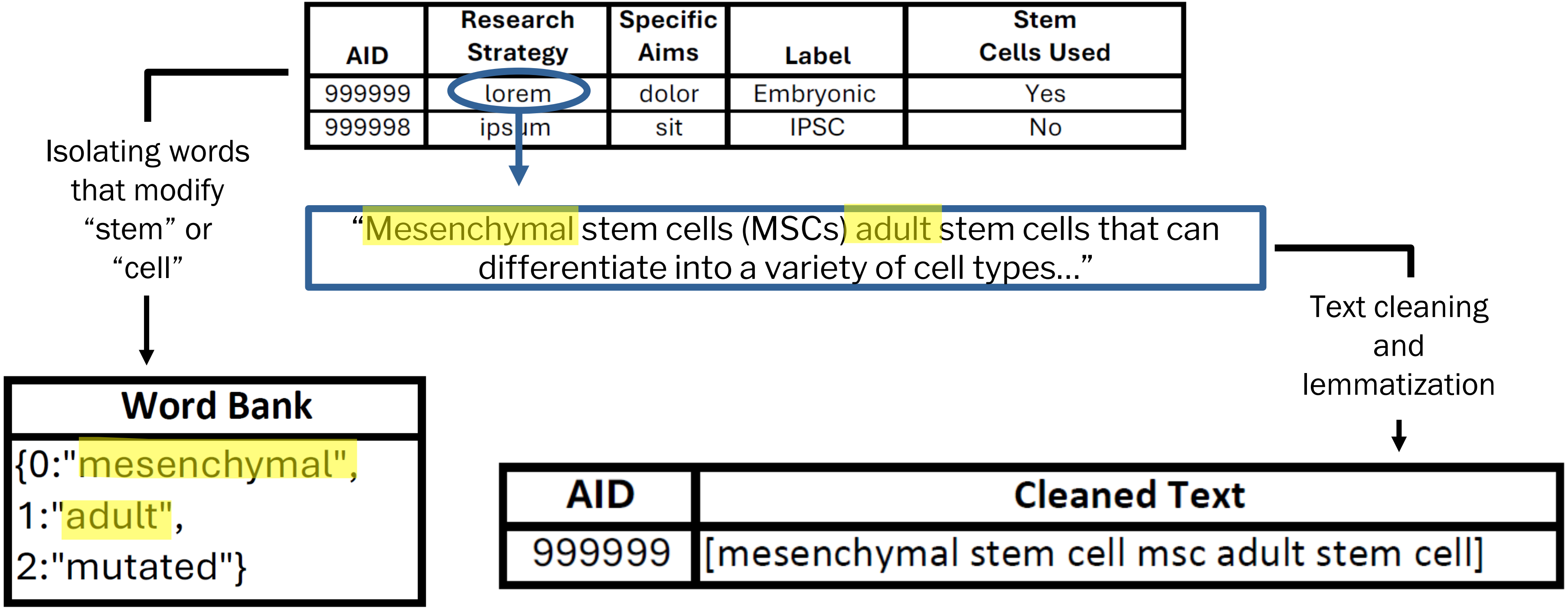


RULE-BASED APPROACH

WORD COUNTS

AID	Research Strategy	Specific Aims	Label	~~Humans Used~~	~~Animals Used~~	Stem Cells Used
999999	lorem	dolor	Embryonic	~~Yes~~	~~Yes~~	Yes
999998	ipsum	sit	IPSC	~~No~~	~~No~~	No

2. WORD COUNTS



AID	Research Strategy	Specific Aims	Label	Stem Cells Used
999999	lorem	dolor	Embryonic	Yes
999998	ipsum	sit	IPSC	No

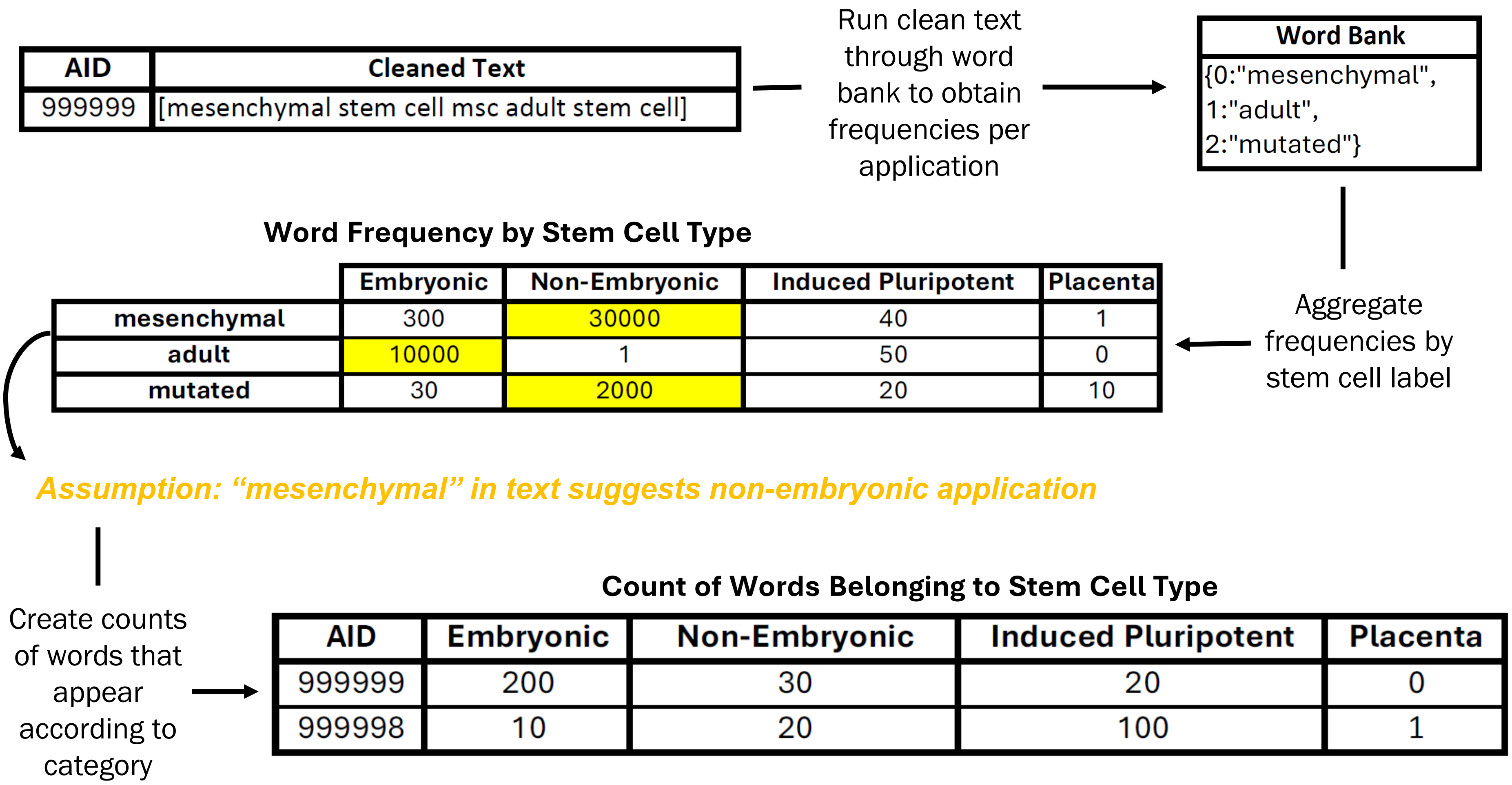
"Mesenchymal stem cells (MSCs) adult stem cells that can differentiate into a variety of cell types…"

Isolating words that modify "stem" or "cell"

Word Bank
{0:"mesenchymal", 1:"adult", 2:"mutated"}

Text cleaning and lemmatization

AID	Cleaned Text
999999	[mesenchymal stem cell msc adult stem cell]



AID	Cleaned Text
999999	[mesenchymal stem cell msc adult stem cell]

Run clean text through word bank to obtain frequencies per application

Word Bank

{0:"mesenchymal",
1:"adult",
2:"mutated"}

Aggregate frequencies by stem cell label

Word Frequency by Stem Cell Type

	Embryonic	Non-Embryonic	Induced Pluripotent	Placenta
mesenchymal	300	30000	40	1
adult	10000	1	50	0
mutated	30	2000	20	10

Assumption: "mesenchymal" in text suggests non-embryonic application

Create counts of words that appear according to category

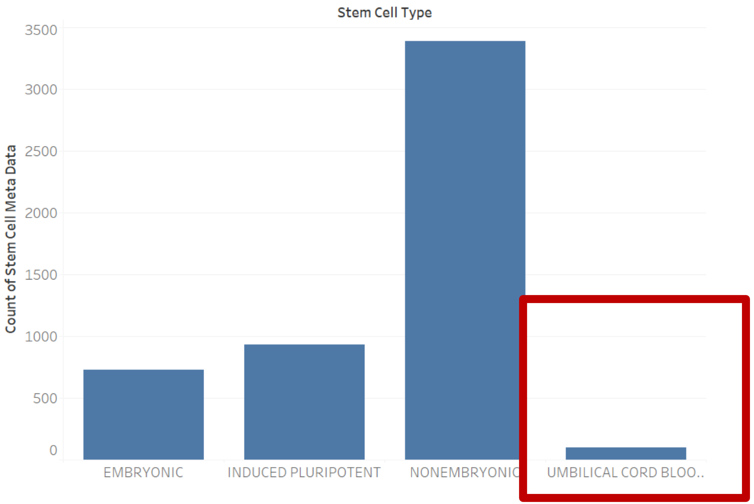
Count of Words Belonging to Stem Cell Type

AID	Embryonic	Non-Embryonic	Induced Pluripotent	Placenta
999999	200	30	20	0
999998	10	20	100	1

SMOTE

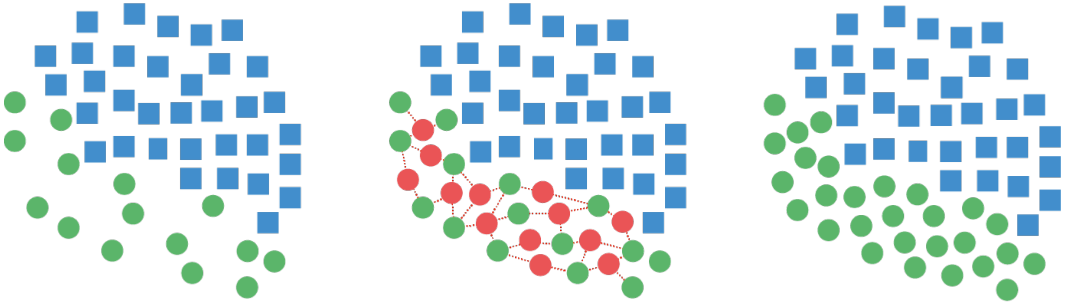
Count of Words Belonging to Stem Cell Type

AID	Embryonic	Non-Embryonic	Induced Pluripotent	Placenta
999999	200	30	20	0
999998	10	20	100	1



SMOTE: Synthetic Oversampling

Synthetic Minority Oversampling Technique



Original Dataset

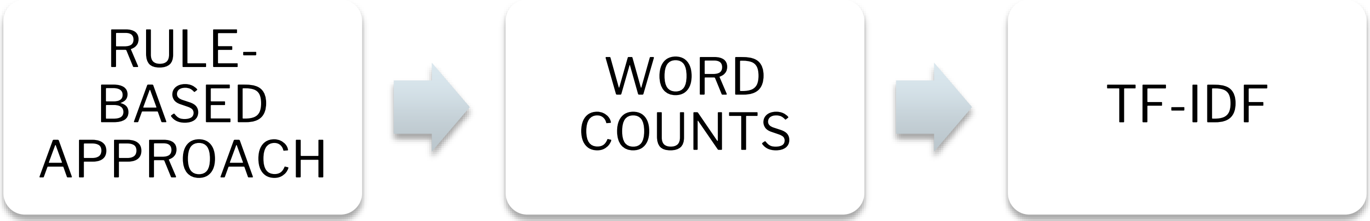
Generating Samples

Resampled Dataset

AID	Placenta
999999	0
999998	1

AID	Placenta
999999	0
----------	0.5
999998	1

3. TFIDF



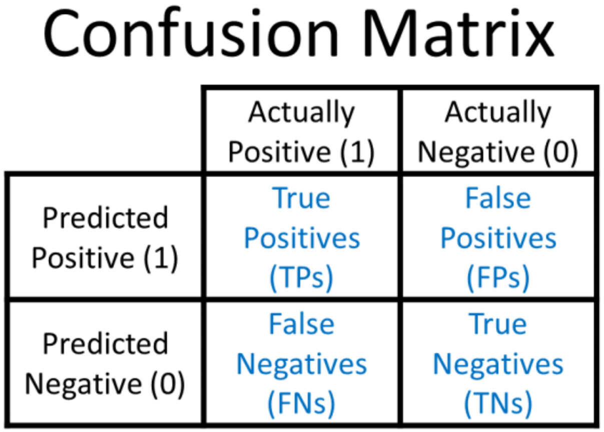
RULE-BASED APPROACH

WORD COUNTS

TF-IDF

CIVIC DIGITAL FELLOWSHIP

SEGWAY: BALANCED ACCURACY



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

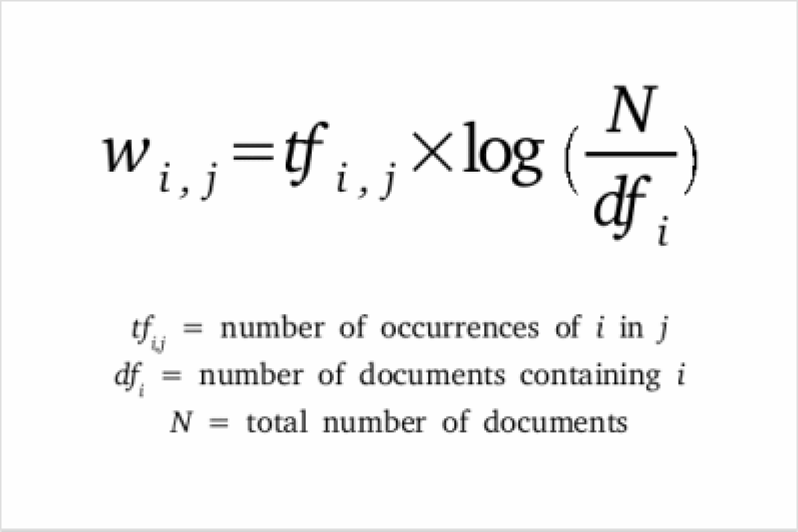
Accuracy – Fraction correctly identified:
(True Positive + True Negative) / Total

Sensitivity – Fraction of Positives correctly identified:
True Positive / (True Positive + False Negative)

Specificity – Fraction of Negatives Correctly Identified:
True Negative / (True Negative + False Positive)

Balanced Accuracy = Average of Sensitivity & Specificity

3. TFIDF



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

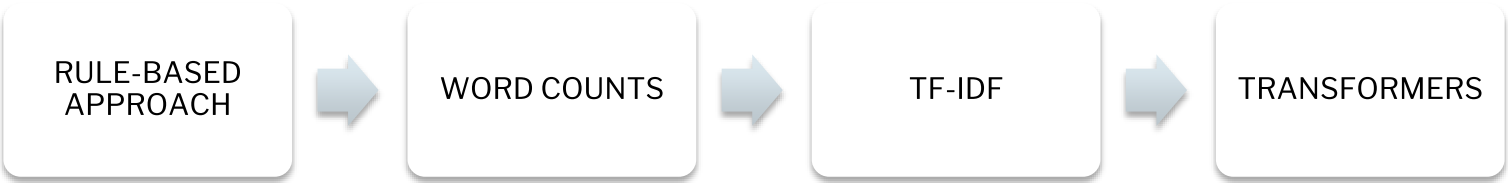
Word Bank
{0:"mesenchymal", 1:"adult", 2:"mutated"}

+

RCDC CATEGORIES
["stemness", "retinal", "progenitor"]

AID	TF-IDF
999999	[["stemness", .30], ["adult", .01], ["progenitor", .15]]
999998	[["stemness", .01], ["adult", .70], ["progenitor", .30]]

4. TRANSFORMERS



RULE-BASED APPROACH

WORD COUNTS

TF-IDF

TRANSFORMERS

Balanced Accuracies:

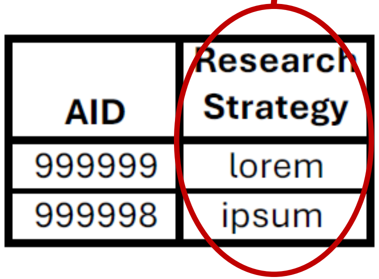
Embryonic – 85.35%

Non-Embryonic – 73.15%

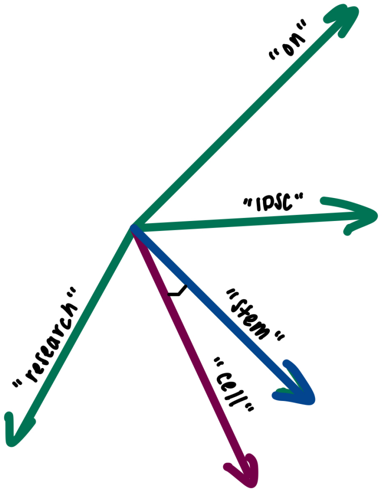
Induced Pluripotent – 69.5%

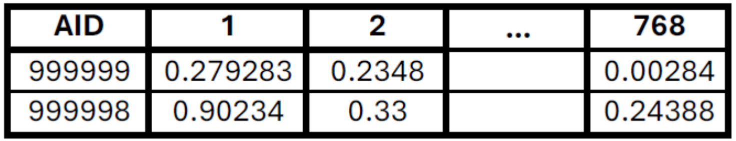
Umbilical – 56.1%

4. TRANSFORMERS



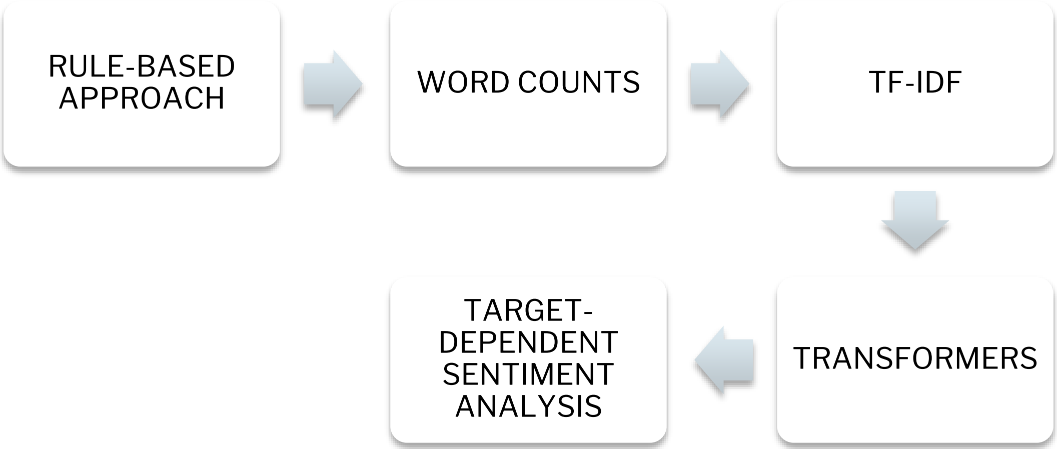
AID	Research Strategy
999999	lorem
999998	ipsum





AID	1	2	...	768
999999	0.279283	0.2348		0.00284
999998	0.90234	0.33		0.24388

4. TRANSFORMERS



Balanced Accuracies:

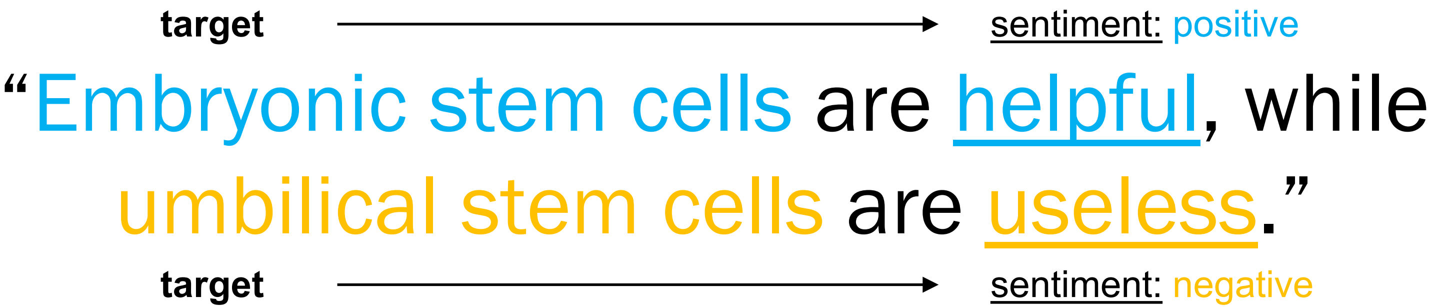
Embryonic – 88%

Non-Embryonic – 70.2%

Induced Pluripotent – 77.2%

CIVIC DIGITAL FELLOWSHIP

5. SENTIMENT ANALYSIS



target → sentiment: positive

"Embryonic stem cells are helpful, while umbilical stem cells are useless."

target → sentiment: negative

THANK YOU!