Deciphering AI from Human Authorship with Machine Learning

Sam Beilenson and Edward Miranda

Key Research Questions

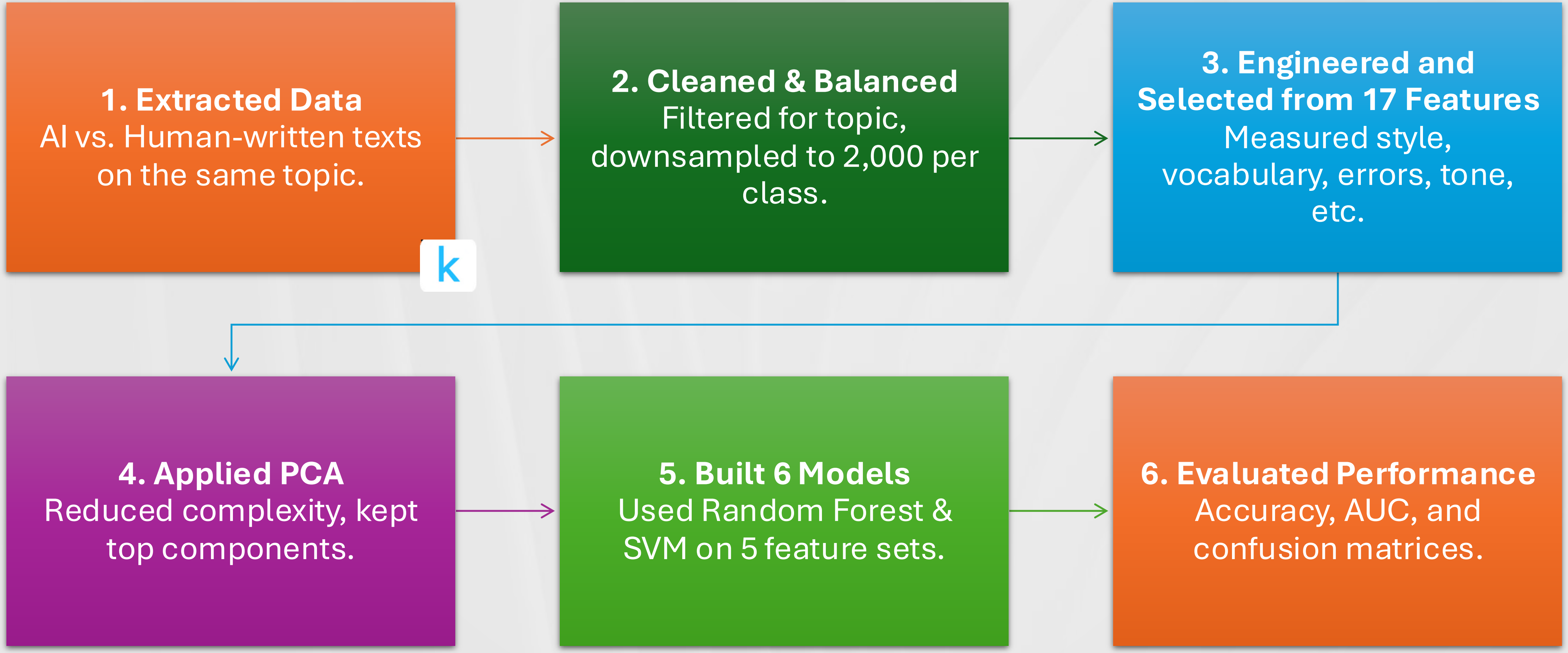
1. Can machines tell AI from human writing?
Are patterns detectable?

2. What features give it away?
Word choice, grammar, sentence style?

3. Does simplifying help?
Can PCA reveal deeper patterns?

4. Is it generalizable?
Would this work on other texts or AI models?

Brief Process Overview



1. Extracted Data
AI vs. Human-written texts on the same topic.

2. Cleaned & Balanced
Filtered for topic, downsampled to 2,000 per class.

3. Engineered and Selected from 17 Features
Measured style, vocabulary, errors, tone, etc.

4. Applied PCA
Reduced complexity, kept top components.

5. Built 6 Models
Used Random Forest & SVM on 5 feature sets.

6. Evaluated Performance
Accuracy, AUC, and confusion matrices.

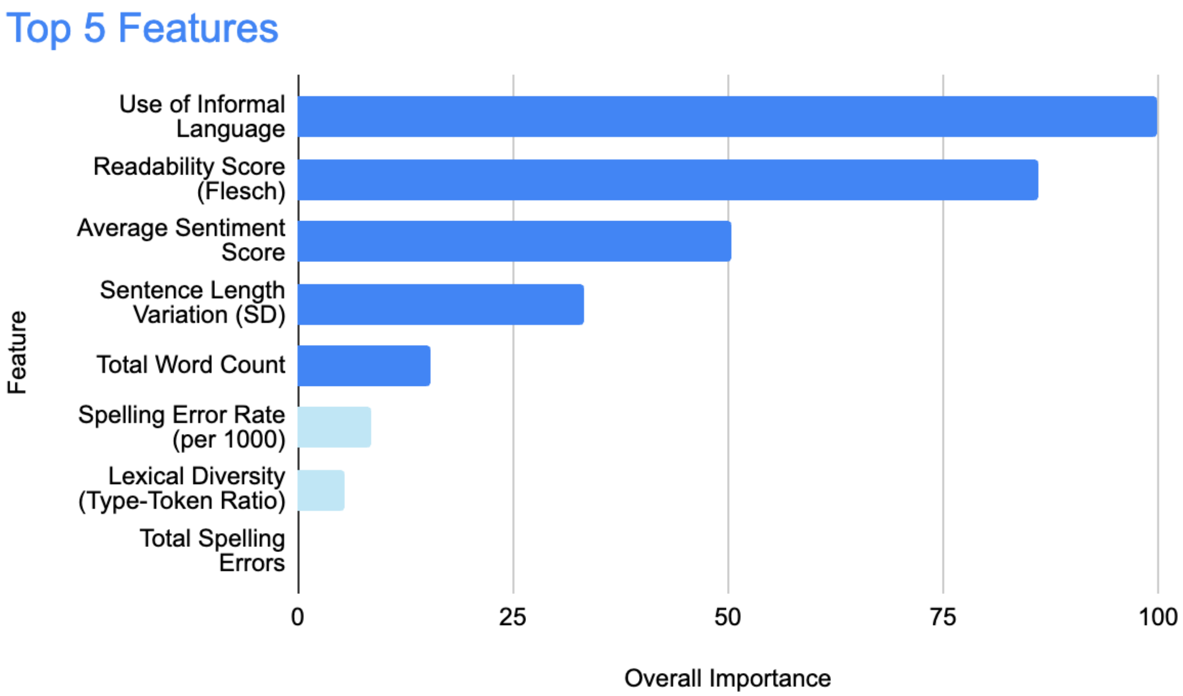
Feature Engineering: Introduction

How features were generated and selected

1. Read sample essays and reflected on common patterns.
2. Drew from personal experience of using AI.
3. Consulted research on detecting AI in **text, images, and audio**.
4. Chose features that felt both humanly intuitive and statistically sound.

How We Ranked Them

- Trained a **Random Forest model** using all engineered features.
- Ranked features by how much they improved predictions.



Informal Language

#	Feature	Test Used	Mean (Human)	Mean (AI)	p-value	Significant?	Notes
1	Informal Language	t-test	5.52	0.47	< 2.2e-16	Yes	Humans used more contractions/slang
2	Readability (Flesch)	t-test	54.82	41.67	< 2.2e-16	Yes	Human text scored as easier to read
3	Sentiment	t-test	0.96	1.44	< 2.2e-16	Yes	AI had more extreme sentiment
4	Sentence Var. (SD)	t-test	10.98	8.63	< 2.2e-16	Yes	Humans varied sentence length more
5	Word Count	t-test	43.49	44.76	< 2.2e-16	Yes	AI used slightly more words on avg

Counted how many slang/contraction words per 1000 words

Readability Flesch

#	Feature	Test Used	Mean (Human)	Mean (AI)	p-value	Significant?	Notes
1	Informal Language	t-test	5.52	0.47	< 2.2e-16	Yes	Humans used more contractions/slang
2	Readability (Flesch)	t-test	54.82	41.67	< 2.2e-16	Yes	Human text scored as easier to read
3	Sentiment	t-test	0.96	1.44	< 2.2e-16	Yes	AI had more extreme sentiment
4	Sentence Var. (SD)	t-test	10.98	8.63	< 2.2e-16	Yes	Humans varied sentence length more
5	Word Count	t-test	43.49	44.76	< 2.2e-16	Yes	AI used slightly more words on avg

Used a built-in readability formula (Flesch) to score each essay

Sentiment (More emotional language??)

#	Feature	Test Used	Mean (Human)	Mean (AI)	p-value	Significant?	Notes
1	Informal Language	t-test	5.52	0.47	< 2.2e-16	Yes	Humans used more contractions/slang
2	Readability (Flesch)	t-test	54.82	41.67	< 2.2e-16	Yes	Human text scored as easier to read
3	Sentiment	t-test	0.96	1.44	< 2.2e-16	Yes	AI had more extreme sentiment
4	Sentence Var. (SD)	t-test	10.98	8.63	< 2.2e-16	Yes	Humans varied sentence length more
5	Word Count	t-test	43.49	44.76	< 2.2e-16	Yes	AI used slightly more words on avg

Averaged emotional word scores from a sentiment dictionary

Sentence Variation

#	Feature	Test Used	Mean (Human)	Mean (AI)	p-value	Significant?	Notes
1	Informal Language	t-test	5.52	0.47	< 2.2e-16	Yes	Humans used more contractions/slang
2	Readability (Flesch)	t-test	54.82	41.67	< 2.2e-16	Yes	Human text scored as easier to read
3	Sentiment	t-test	0.96	1.44	< 2.2e-16	Yes	AI had more extreme sentiment
4	Sentence Var. (SD)	t-test	10.98	8.63	< 2.2e-16	Yes	Humans varied sentence length more
5	Word Count	t-test	43.49	44.76	< 2.2e-16	Yes	AI used slightly more words on avg

Measured how much sentence length varied in each essay

Word Count

#	Feature	Test Used	Mean (Human)	Mean (AI)	p-value	Significant?	Notes
1	Informal Language	t-test	5.52	0.47	< 2.2e-16	Yes	Humans used more contractions/slang
2	Readability (Flesch)	t-test	54.82	41.67	< 2.2e-16	Yes	Human text scored as easier to read
3	Sentiment	t-test	0.96	1.44	< 2.2e-16	Yes	AI had more extreme sentiment
4	Sentence Var. (SD)	t-test	10.98	8.63	< 2.2e-16	Yes	Humans varied sentence length more
5	Word Count	t-test	43.49	44.76	< 2.2e-16	Yes	AI used slightly more words on avg

Counted total number of words in each essay

Note: Word count may not be the best feature in future models — it's often influenced by writing prompts and doesn't reflect writing *style* as clearly as the other features.

Top 5 features

Informal Language

Humans use contractions and slang—AI rarely does.

Readability (Flesch Score)

AI text is often less readable than natural human writing.

Sentiment Score

Human writing has less emotional language.

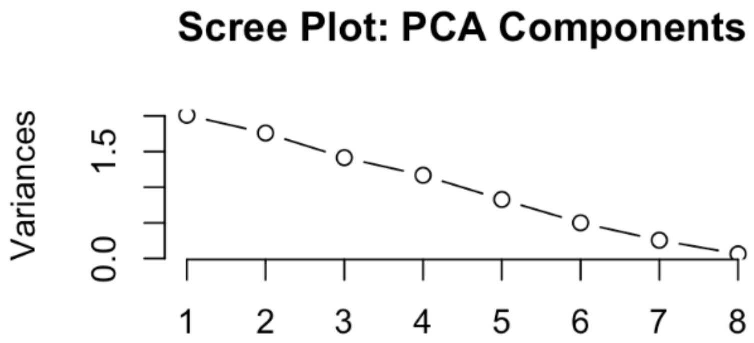
Sentence Length Variation

Humans mix short and long sentences—AI keeps it uniform.

Total Word Count

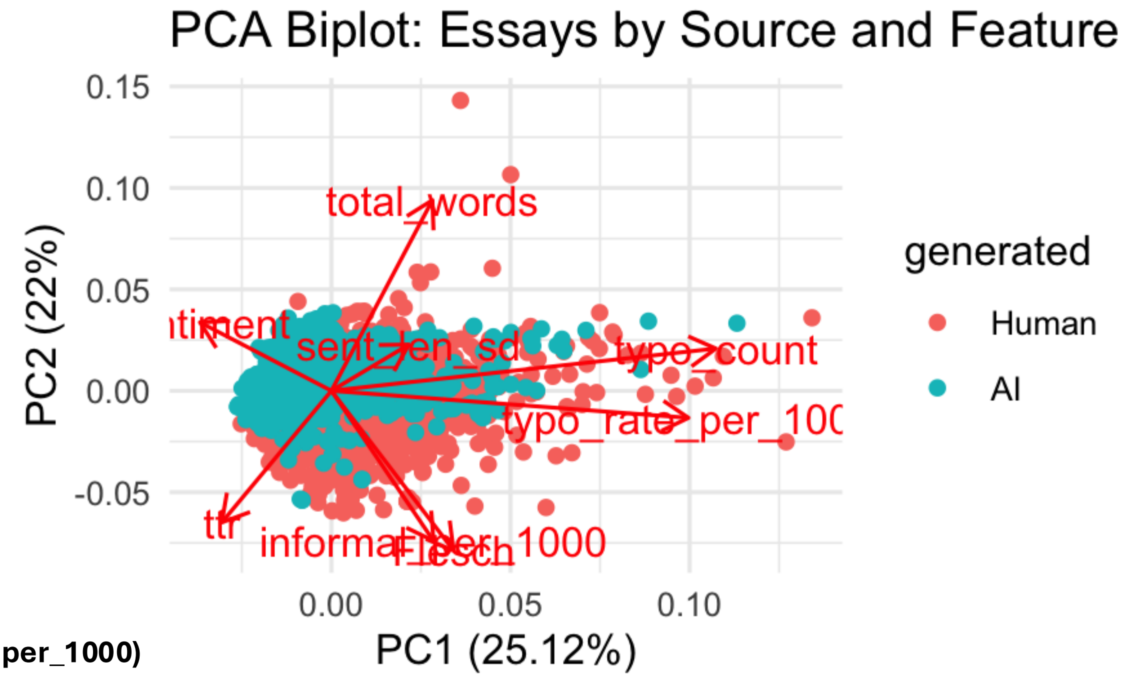
AI essays had a higher word count *in this dataset*.

Principle Component Analysis



Scree Plot: PCA Components

- Each dot is an essay — red = Human, blue = AI.
- PCA groups similar essays together based on patterns.
- The arrows show which features (like sentence length or typos) influence those groupings.
- Human and AI essays often fall in different areas — meaning their writing has detectable patterns.

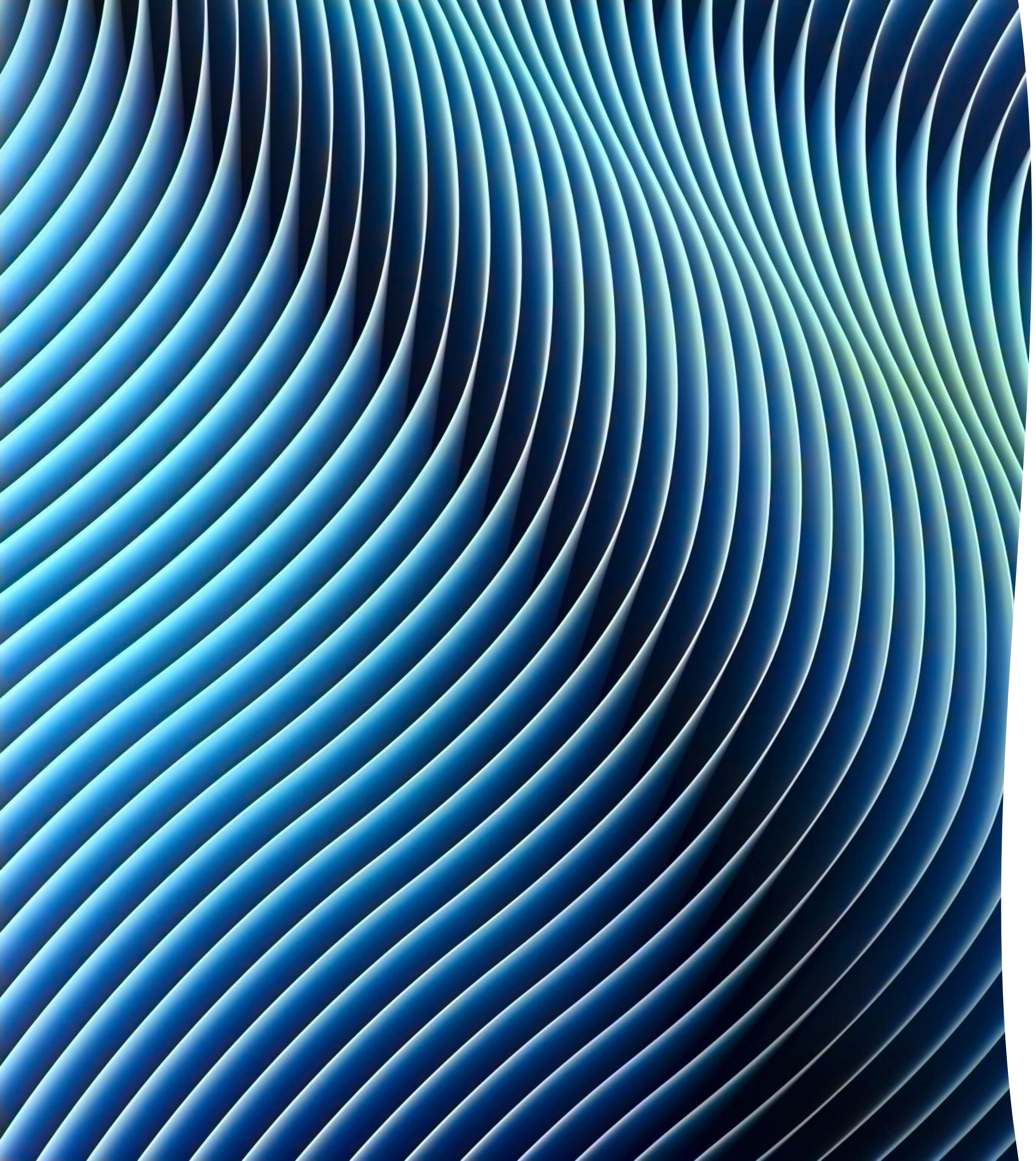


PCA Biplot: Essays by Source and Feature

Strongest arrows on this plot:

1. **Total Words**
2. **Informal Language (informal_per_1000)**
3. **Flesch Readability Score**
4. **Typo Rate (typo_rate_per_1000)**

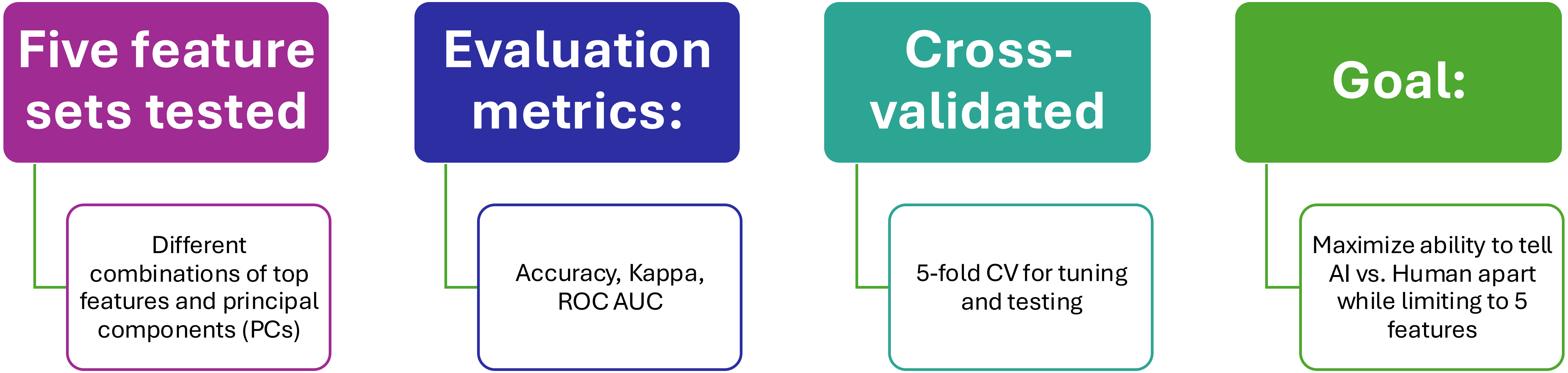
Model Building Overview



Building The Classifier

☑	**Data Prep:**	• Balanced sample — 2,000 Human and 2,000 AI essays
☑	**Feature Engineering:**	• Selected top 5 strongest features based on feature importance and research
☑	**Dimensionality Reduction:**	• PCA to capture patterns in the data
	Train/Test Split:	• 80% Training / 20% Testing
	Modeling Algorithms:	• Random Forest and SVM (Support Vector Machine)

Model Strategy



Five feature sets tested

Different combinations of top features and principal components (PCs)

Evaluation metrics:

Accuracy, Kappa, ROC AUC

Cross-validated

5-fold CV for tuning and testing

Goal:

Maximize ability to tell AI vs. Human apart while limiting to 5 features

Hyperparameters and Cross-Validation

- **Hyperparameters** like the number of features considered at each split (mtry) and minimum samples per leaf (min_n) were tuned to maximize model performance.
- **Random search** was used to explore 10 random combinations during tuning.
- **5-fold Cross-Validation** was applied:
  - The training data was split into 5 parts.
  - The model trained on 4 parts and validated on the 5th, rotating folds.
  - This helps ensure the model generalizes well and reduces overfitting.

Why Random Forest and SVM?

Random Forest:

- Handles **text-derived features** (like readability, sentiment) that can be **noisy or correlated**.
- **Feature importance** scores help **identify key differences** between AI and human writing.
- Works **reliably even with a smaller sample size** (n = 4000 essays).

Support Vector Machine (SVM):

- Good at **finding subtle differences** when the classes (AI vs Human) are **close together**.
- Excels in **high-dimensional spaces**, which is common with **engineered language features**.
- Provides a **robust comparison** to Random Forest performance.

Model Performance

Rank	Model #	Feature Set Name	Accuracy	ROC AUC	Kappa
1	**1**	**Engineered Only**	**0.9225**	**0.97895**	**0.845**
2	2	2 PCs + 3 Engineered	0.91875	0.97434	0.8375
3	5	1 PC + 4 Engineered	0.92	0.97209	0.84
4	6	SVM – Top 5 Engineered	0.92	0.97174	0.84
5	3	Top 5 PCs Only	0.91375	0.97214	0.8275
6	4	4 PCs + 1 Engineered	0.90875	0.96755	0.8175

We believe ROC AUC is better than accuracy or kappa here because it measures how well the model separates AI from Human across all decision thresholds, not just whether predictions are "right" or "wrong" at one cutoff.

Test set
Top Model Summary

	Predicted: Human	Predicted: AI
Actual: Human	**376**	**38**
Actual: AI	**24**	**362**

62 total errors:

- **38 Human essays** incorrectly predicted as AI
- **24 AI essays** incorrectly predicted as Human

Limitations/Generalizability Concerns

Control for Simpler Features

- Future work should better control for **word count, contractions**, and **spelling errors** to isolate deeper linguistic signals of authorship.

Find a way to engineer the top features according to research:

- Perplexity
- Burstiness
- Syntactic Diversity
- Semantic Coherence
- Lexical Diversity
- Stylistic Variation
- Minor Errors (or Lack Thereof)

Sample Size

Our smaller sample (n = 5000) was a tradeoff for performance — scaling up would help generalize findings.

Why It Matters

Understanding AI authorship is **ethically urgent**:

- AI text can **bypass learning** and diminish education.
- Fuels **deepfakes** and **ransom tactics**
- Risks **misinformation** and **polarization**