Title & Objective

Legal Document Classification

Document TextCategory

"The court rules in favor of the defendant “ . "Judgment"
This agreement is valid for one year. "Contract"
Take notice of the change. "Notice”

Data Processing Cleaning

Original ;

"The court rules in favor of the defendant."

1. Lowercase ;

"the court rules in favor of the defendant."
Example : Makes "Court" and "court" the same word.

2. Stopword Removal ;

Removes common words like *the*, *in*, *of*, *a* that don't define the category.

3. Lemmatization ;

Reduce rules to itS rule

Feature Extraction (TF-IDF)

Machine learning models only understand numbers. TF-IDF converts the cleaned text into numerical vectors, assigning a score to each word based on its importance

TF (Term Frequency): How often a word appears in the current document.

IDF (Inverse Document Frequency): Penalizes common words that appear across *all* documents.

Word	**Document 1 (judgement)**	**Document 2 (contract)**	**Document 3 (notice)**	**TF - IDF SCORE**
COURT	**1**	**0**	**0**	**High (Specific to Judgment)**
AGGREMENT	**0**	**1**	**0**	**High (Specific to Contract)**
NOTICE	**0**	**0**	**1**	**High (Specific to Notice)**
VAILD	**0**	**1**	**0**	**Medium**

The model sees each document as a list of numbers (a vector):

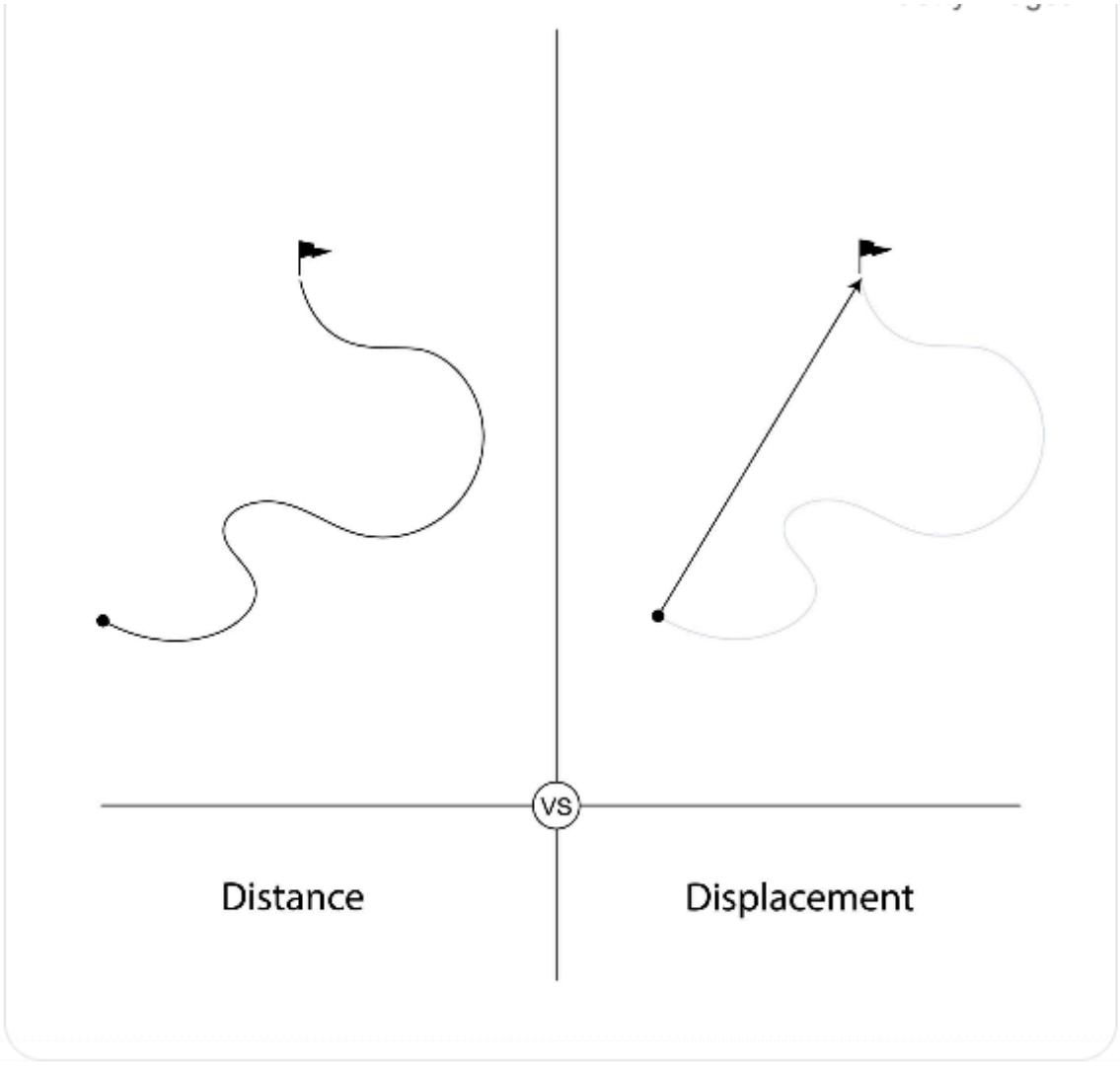
- Judgment Vector: [0.8,0.0,0.0,0.0]
- Contract Vector: [0.0,0.9,0.0,0.7]
- Notice Vector: [0.0,0.0,0.9,0.0]

4. Model Training (SVM)

The Support Vector Machine (SVM) model takes these numerical vectors and plots them in a high-dimensional space.

- It finds the best boundary (hyperplane) to separate the 'Judgment' points from the 'Contract' points, and so on.

- After training, this hyperplane is fixed and used to classify new points (new documents).



Precision, Recall, and F1-Score (Detailed Metrics)

These metrics are calculated *per category* and provide a deeper understanding than just accuracy:

Metric	Simple Explanation	Calculation based on our error
Precision	**How trustworthy is the prediction?** Out of all documents predicted as "Contract," how many were *actually* "Contract"?	Predicted ContractTrue Contract=21 (50%)
Recall	**How many did we find?** Out of all documents that *were* "Notice," how many did the model correctly find?	Actual NoticeTrue Notice=10 (0%)
F1-Score	**The single summary score.** A balance of Precision and Recall.	Lower F1 for 'Notice' shows poor performance on that class.

Confusion Matrix (Graph)

The Confusion Matrix is a visual table that shows where the model is confused (where the prediction and actual category differ).

Actual / Predicted	Contract	Judgment	Notice
Contract	1 (Correct)	0	0
Judgment	0	1 (Correct)	0
Notice	1 (Error!)	0	0 (Wrong!)