National PhD in Artificial Intelligence / AI for Society

Statistical Learning and Large Data (SLLD) 1 & 2 - Prof. Francesca Chiaromonte

Will Your Paper Get Accepted?

Predicting NeurIPS Submission Success

Students: **Lorenzo Emer & Riccardo Porcedda**

May 9, 2025

Context: Predicting paper acceptance can help authors better understand the factors influencing success, assist reviewers and organizers in streamlining the evaluation process, and potentially uncover hidden **biases** or **patterns** that shape scientific careers and knowledge dissemination [1, 2, 3, 5].

Goal: Predict the acceptance of papers submitted to NeurIPS, one of the top AI conferences, using only information available at submission (i.e., the manuscript and metadata).

RQ1 : Can we accurately predict the acceptance of NeurIPS submissions using features derived from the paper text and structure?

RQ2 : Which textual or structural features are most influential in determining the acceptance of a NeurIPS submission?

Data Collection & Preprocessing

Scraping from **NeurIPS 2024** on OpenReview.
Different **Python libraries** employed: **Selenium** (browser automation and interaction), **BeautifulSoup** (HTML parsing and extraction), **Pandas**, **NumPy** (storing and handling extracted data)

Text extraction from the PDFs (PyPDF2 on Python), **preprocessing** (removal of "NeurIPS Paper Checklist", appendices, and references sections to isolate the main content).

Text Representation: construction of a **TF-IDF Matrix** (Term Frequency - Inverse Document Frequency), after text **tokenization**.

Data Description

4238 papers

4037 accepted

201 rejected

Severe Class Imbalance!

Features

- **TD-IDF features (27756)**
- **Textual features:**
  - Word count
  - Unique word count
  - Average sentence length
- **Structural features:**
  - Number of figures
  - Number of tables
  - Number of detected equations
  - Number of references
- **Metadata features:**
  - Title length
  - Abstract length
  - Supplementary Material (BOOL)

TARGET VARIABLE

ACCEPT (BOOL)

Dimensionality Reduction

We apply **Singular Value Decomposition (SVD)**, which can be considered as a generalization of PCA for non-square matrices, like the TF-IDF matrix. That is what makes it a standard choice in NLP pipelines.

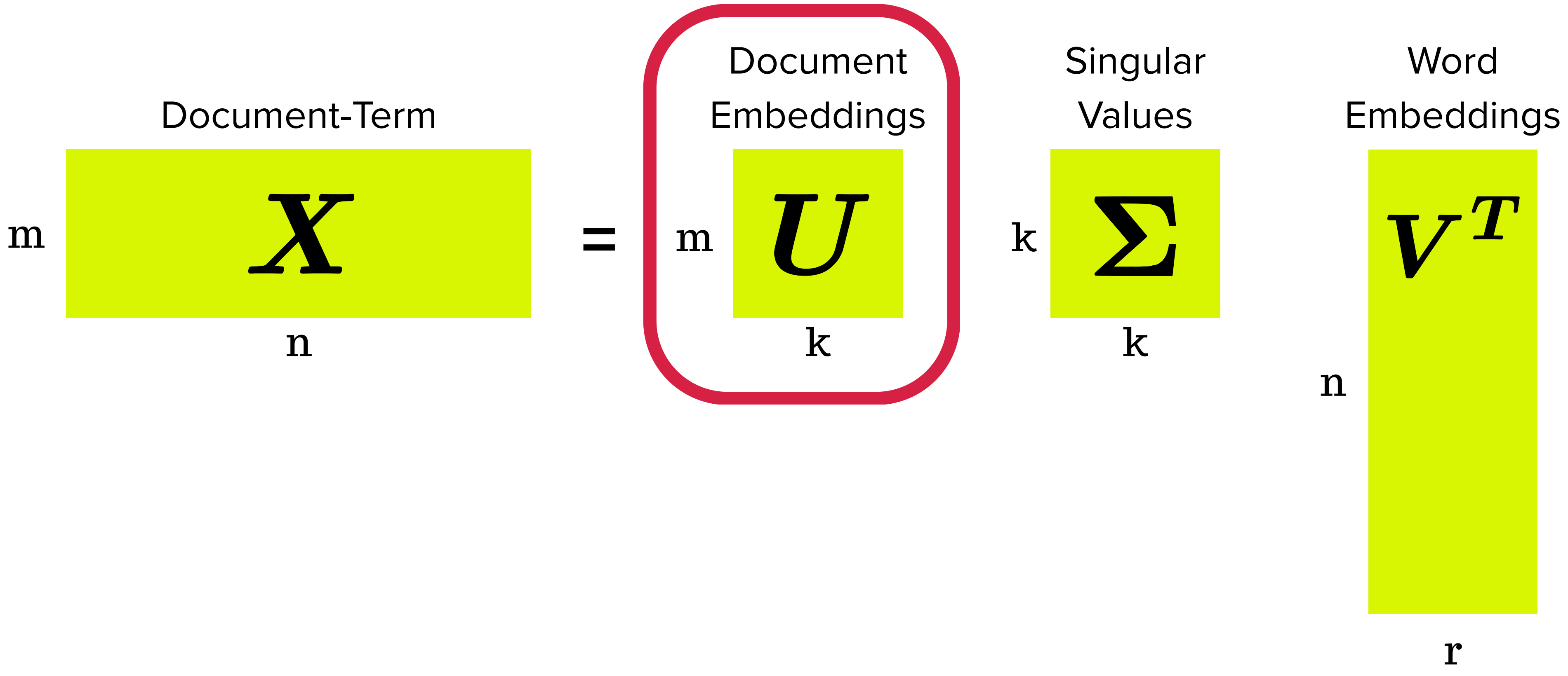
SVD decomposes any matrix into three components:

$$\boldsymbol{M} = \boldsymbol{U\Sigma V^T} \qquad\qquad \boldsymbol{M} \simeq \boldsymbol{U_k \Sigma_k V_k^T}$$

- **U** and **V** are semiunitary matrices.
- **Σ** is a square diagonal matrix containing the singular values.

Dimensionality Reduction: By retaining only the top-k singular values and corresponding vectors, SVD provides a low-rank approximation of the original matrix [6]

SVD Interpretation & PCA



PCA is often computed using SVD of the data matrix (centered by subtracting the mean)

$$C = \frac{1}{n-1} X X^T \qquad X = U \Sigma V^T \longrightarrow C = V \frac{\Sigma^2}{n-1} V^T$$

Classification Models

Using the reduced feature set, we trained several models to predict paper acceptance (test set 30%) [7]

Logistic Regression (with LASSO) - BASELINE MODEL
A simple, interpretable linear model; used L1 regularization to select the most predictive features [4].
Decision Tree:
A rule-based model that splits data into decision paths
Random Forest:
An ensemble of decision trees that reduces variance; often powerful, but less interpretable.
Support Vector Machine (SVM):
A margin-based classifier that finds the optimal separating hyperplane; performs well in high-dimensional spaces.
k-Nearest Neighbors (k-NN):
A non-parametric model that predicts based on similarity to nearby data points; sensitive to scaling and feature density.

Variants of Logistic Regression

To enhance performance and interpretability, we tested multiple logistic regression strategies:

LASSO Feature Selection:
Applied L1 regularization to identify a sparse set of informative features.
Class-Weighted Logistic Regression:
Applied class weights to handle label imbalance (more accepted than rejected papers).
SIS + LASSO (Sure Independence Screening):
Pre-selected features based on marginal correlation before applying LASSO.
SIS Only (no LASSO):
Trained directly on SIS-selected features.
Stability-Based Downsampling:
Kept only variables that were stable (selected in >20% of 20 subsamples).

Model Evaluation

Confusion Matrix: Shows True Positives, False Positives, etc., for understanding prediction distribution.

Precision: Proportion of predicted accepted papers that were actually accepted.
➜ How many predicted positives were correct?

Recall: Proportion of actually accepted papers correctly predicted.
➜ How many actual positives did we catch?

F1-Score: Harmonic mean of precision and recall — balances both aspects.

AUC (Area Under the Curve): Measures ranking quality; how well the model separates accepted vs. rejected papers.

Average Precision (AP): A more informative metric than accuracy in imbalanced settings — captures ranking performance.

Results

- **Best overall model: Weighted Logistic Regression — combines strong performance (F1 = 0.90, AUC = 0.81) with interpretability and feature sparsity.**
- Other models: SVM: F1 = 0.97, AUC = 0.80 — excellent recall but less interpretable.
- k-NN: F1 = 0.97, but lower AUC (0.61), similar to SVM.
- Random Forest & Decision Tree: Extremely high F1 and recall, but **predicted nearly everything as accepted — low AUC, limited real-world usefulness.**

Feature selection strategies:

- LASSO and retraining provided sparsity and decent performance (F1 ~0.83–0.84), useful for explainability.
- SIS alone gave high recall and F1, but with poorer AUC — indicating weak discrimination power.
- SIS + LASSO failed in practice (underfit or unstable)

A few confusion matrices...

LOGIT + LASSO

	Predicted	
Actual	0	1
0	32	29
1	337	873

WEIGHTED LOGIT

	Predicted	
Actual	0	1
0	41	20
1	207	1003

SIS (NO LASSO)

	Predicted	
Actual	0	1
0	12	49
1	129	1081

SVM

	Predicted	
Actual	0	1
0	1	60
1	4	1206

Going back to the features

Features Kept by LASSO (logistic regression)

Feature	Coefficient
`word_count`	+0.0009189
`unique_word_count`	–0.006665
`num_tables`	+0.1666
`num_figures`	+0.2563
`avg_sentence_length`	+0.04305

+ 48 SVD dimensions

Reversing SVD

highest BETA	lowest BETA	highest p-value	lowest p-value
ace	simplex	handle_case	graph_generation
actionvalue_function	free	type_error	llava
dynamic_scene	cat	open_vocabulary	separation
graph_generation	scan	initialize_parameter	blurring
blurring	reaction	parallel_computing	delay
llava	probability_mass	latent_dynamic	free
overlap	separation	independent_run	dynamic_scene
pip	sequential	uncertain	sequential
emp	ope	cooccurrence	amplitude
request	amplitude / volunteer	phase	ope / volunteer

Discussion

- **Weighted Logistic Regression** shows the best balance: strong performance (F1 = 0.90, AUC = 0.81) and interpretable features
- **Positive signals**: number of tables and figures (reviewers favor clarity and evidence presentation), average sentence length (moderate impact)
- **Negative signal:** unique_word_count, possibly penalizing diverse or unfocused writing writing
- **Word-level:** presence of cutting-edge ML topics (llava, graph_generation, actionvalue_function) — showing that specific domains are more likely to be accepted.

Next Steps

- Retrieve **more data**: other ML conferences, other NeurIPS years to improve result generalization
- Test **more advanced models** (e.g., DL-based or gradient boosting models)
- Expand with **network-based** analyses (authors / affiliations)

Sources

[1] D. J. Joshi, A. Kulkarni, R. Pande, I. Kulkarni, S. Patil, and N. Saini. Conference paper acceptance prediction: Using machine learning. In A. Dutta, L. C. Jain, and S. Aggarwal, editors, Machine Learning and Information Processing: Proceedings of ICMLIP 2020, pages 143–152. Springer Singapore, 2021.

[2] A. C. Ribeiro, A. Sizo, H. Lopes Cardoso, andL. P. Reis. Acceptance decision prediction in peer-review through sentiment analysis. In G. Mar-reiros, F. S. Melo, N. Lau, H. Lopes Cardoso, and L. P. Reis, editors, Progress in Artificial Intelligence. EPIA 2021, volume 12981 of Lecture Notes in Computer Science, pages 757–769. Springer, Cham, 2021.

[3] M. Skorikov and S. Momen. Machine learning approach to predicting the acceptance of academic papers. In 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), pages 113–117. IEEE, 2020.

[4] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal StatisticalSociety. Series B (Methodological), 58(1):267–288, 1996.

[5] W. Wang, J. Zhang, F. Zhou, P. Chen, and B. Wang. Paper acceptance prediction at the institutional level based on the combination of individual and network features. Scientometrics, 126:1581–1597, 2021.

[6] Lange, K. (2010). Singular Value Decomposition. In: Numerical Analysis for Statisticians. Statistics and Computing. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-5945-4_9

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, No. 1). New York: springer.

Q&A

Thank you for your attention!