UNIVERSITY of VIRGINIA | SCHOOL of DATA SCIENCE

Optimizing the ALMA Research Proposal Process with Machine Learning

Brendan Puglisi, Arnav Boppudi, Kaleigh O’Hara, Noah McIntire, Ryan Lipps

May 1, 2024

Table of Contents

- About ALMA
- Project Significance
- Data Discussion
- Model
- Limitations & Assumptions
- Conclusions

ALMA Observatory

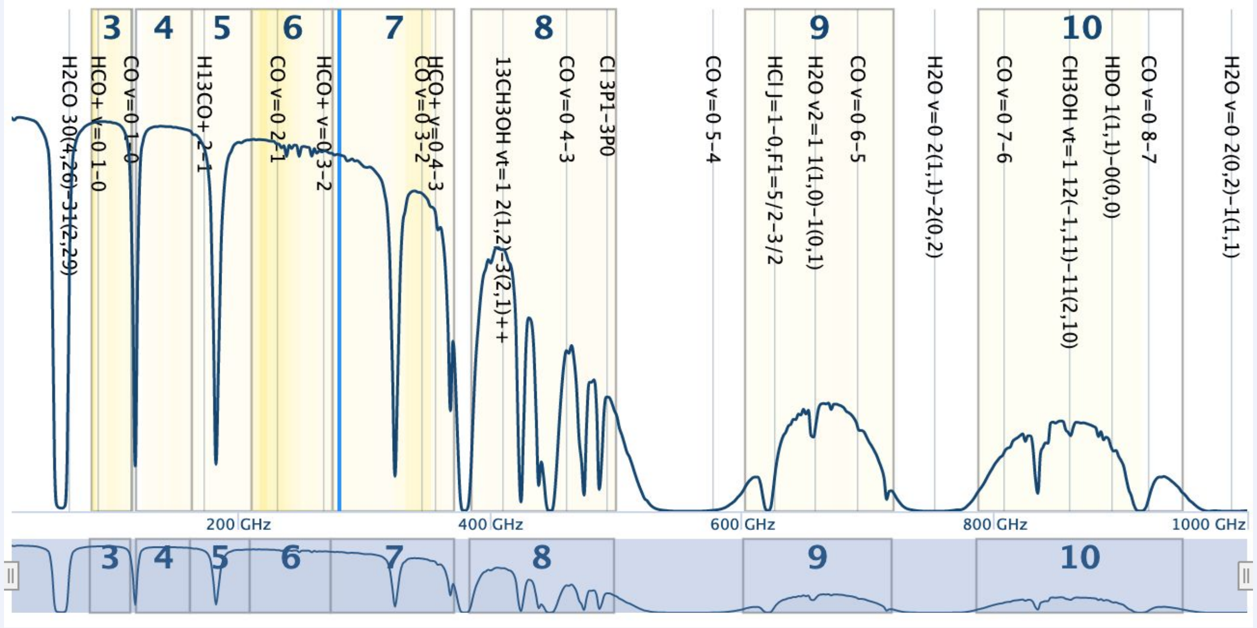
- The ALMA Observatory is located in the Atacama Desert in northern Chile
- The state-of-the-art radio telescope array consists of 66 high-precision antennas that observe electromagnetic radiation outside of visible light

Project Significance

- ALMA is open to anyone to use, based on a proposal process
- Reduce the time and effort required for researchers to prepare proposals
- Our work will help the astronomy community by simplifying the technical aspects of proposal writing, leading to more precise and effective observations

Data Discussion

- Two types of projects:
  - Spectral Line and Continuum
- Line projects require specific measurement setup



Spectral Line Measurements

Target Variable

Project Title and Abstract

Predictor Variable

4,586

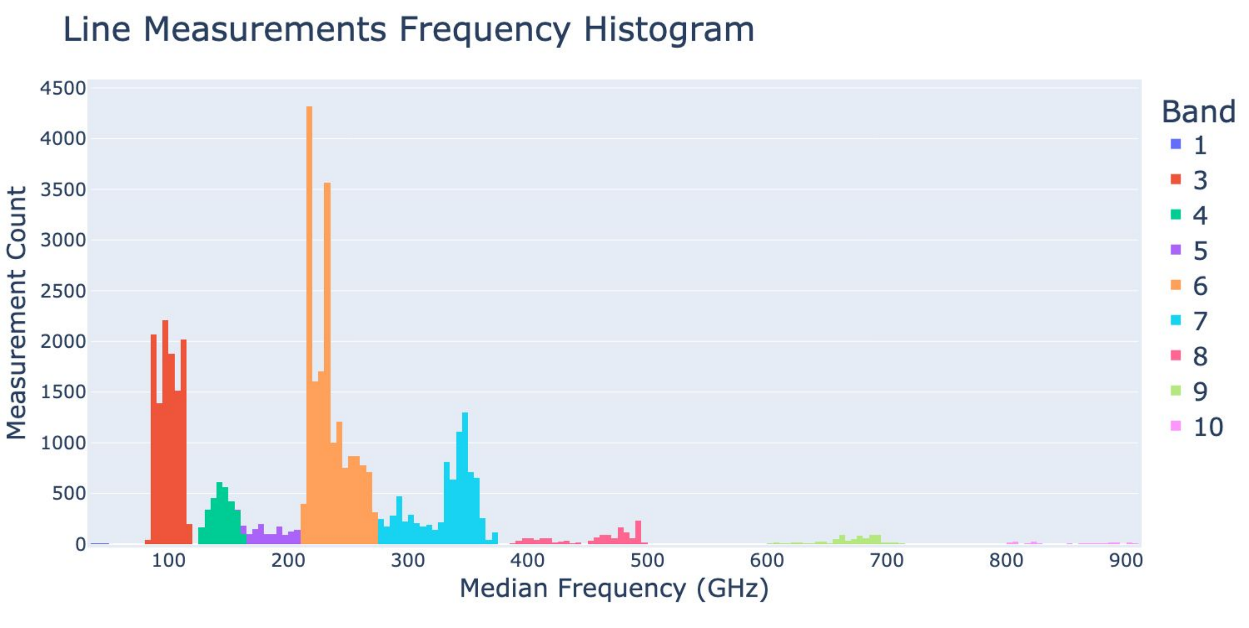
Previous Projects

67,439

Total Measurements

Data Discussion Cont.

- Distribution of measurements across band shows that the vast majority exist in band 3 and 6 and in the lower frequency ranges.



3,628

Line Projects

75%

Have fewer than 13 measurements

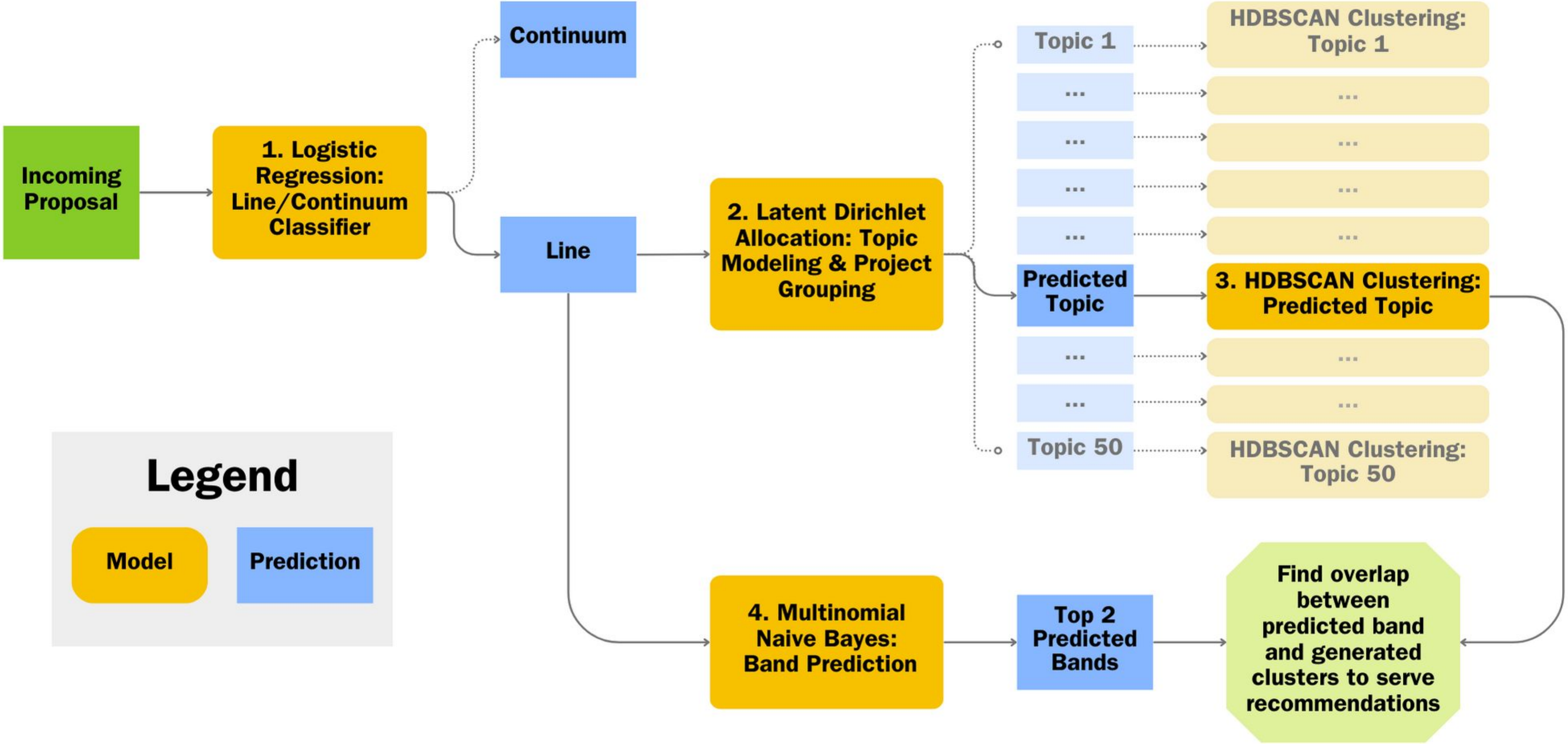
450

Outlier projects have more than 26 measurements

82%

Have measurements in only one frequency band

Model Prediction Pipeline



Incoming Proposal

1. Logistic Regression: Line/Continuum Classifier

Continuum

Line

2. Latent Dirichlet Allocation: Topic Modeling & Project Grouping

Topic 1

...

...

...

...

Predicted Topic

...

...

Topic 50

HDBSCAN Clustering: Topic 1

...

...

...

...

3. HDBSCAN Clustering: Predicted Topic

...

...

HDBSCAN Clustering: Topic 50

4. Multinomial Naive Bayes: Band Prediction

Top 2 Predicted Bands

Find overlap between predicted band and generated clusters to serve recommendations

Legend

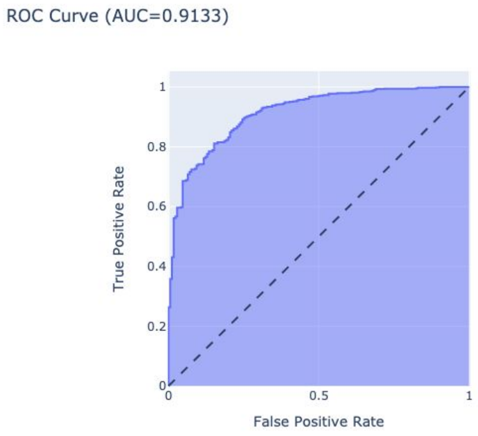
Model

Prediction

01 Project Classification: Logistic Regression

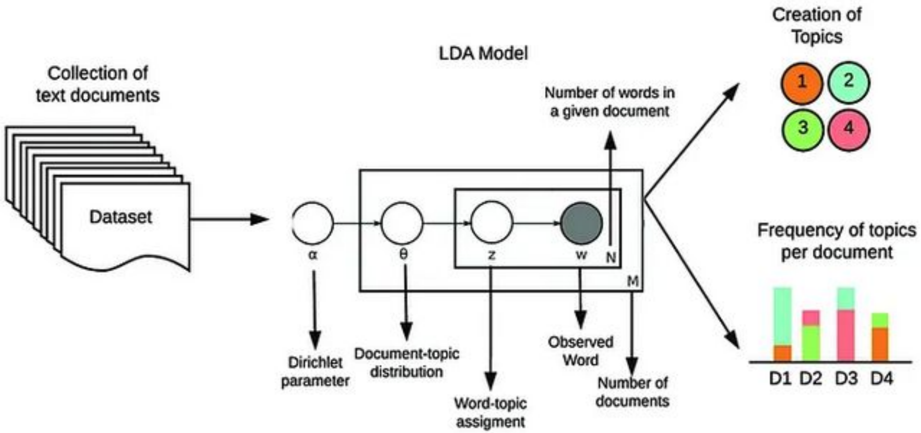
- Vectorize title and abstract using TF-IDF
- Vector used as features to classify a project as either line or continuum
- Only projects with line observations are of interest to us
- Accuracy of 90.02%
- Correctly predicted line 96.41% of the time
- Correctly predicted continuum 59.42% of the time

	Predicted Continuum	Predicted Line
True Continuum	104	71
True Line	26	699



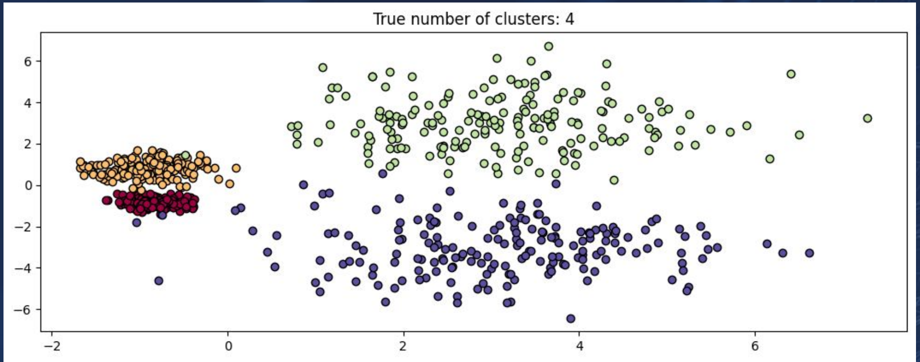
02 Project Grouping: LDA

- Generate 50 topics
- Group projects into topics generated by LDA
- Projects are assigned to their “max topic”
- Topic 25 most heavily weighted words: *bar, gmcs, molecular, spiral, galaxy*
- Topic 37 most heavily weighted words: *mass, chemical, chain, protostars, wccc*

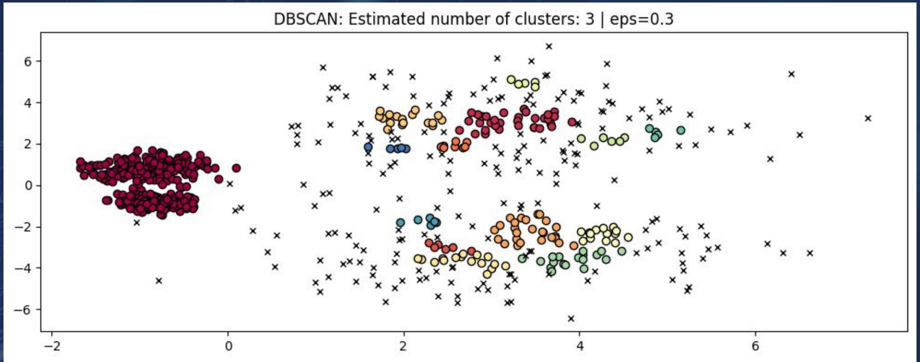


03 Measurement Clustering: HDBSCAN

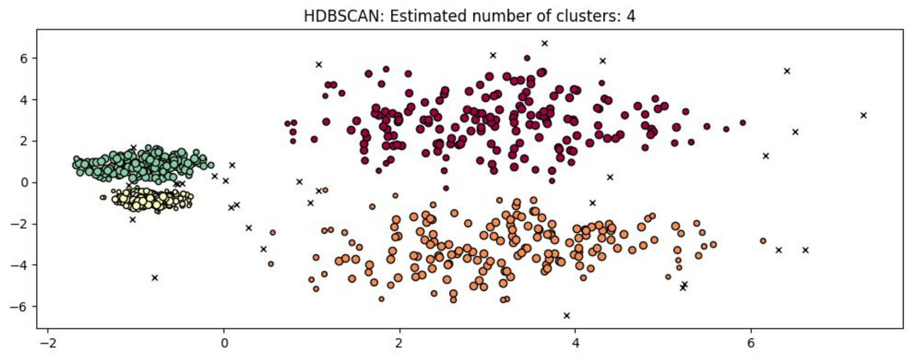
- Cluster measurements for projects within each LDA topic
- Accounts for different densities in measurement distribution
    - Keeps clusters representative of overall measurement distribution
- Ensure clusters are not too large
- Average noise 14.59 ± 0.05%
- Score clusters with count of measurements and projects



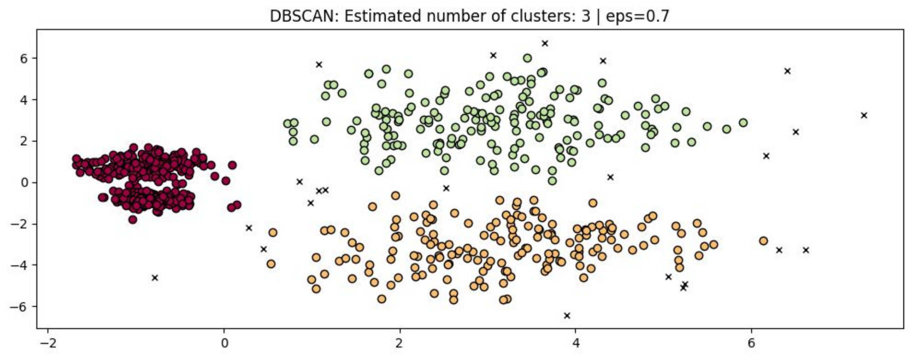
True number of clusters: 4



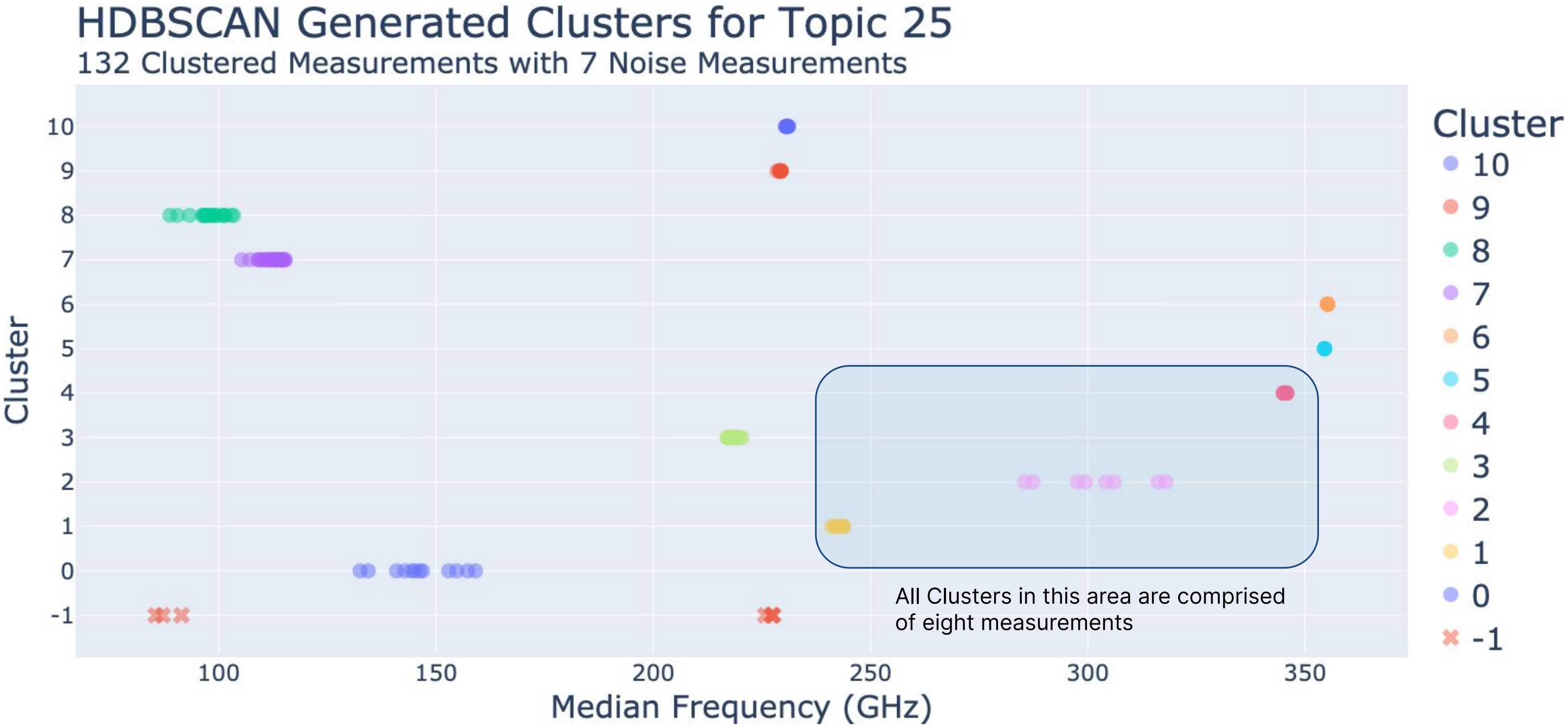
DBSCAN: Estimated number of clusters: 3 | eps=0.3



HDBSCAN: Estimated number of clusters: 4



DBSCAN: Estimated number of clusters: 3 | eps=0.7


HDBSCAN Generated Clusters for Topic 25

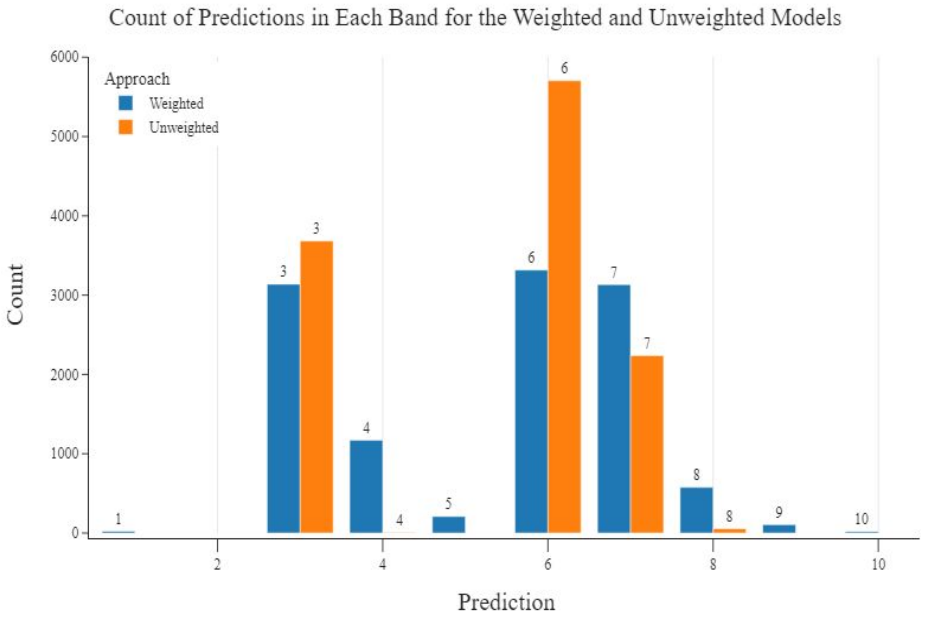
132 Clustered Measurements with 7 Noise Measurements

All Clusters in this area are comprised of eight measurements

04 Band Prediction: Multinomial Naïve Bayes

- Text Preprocessing:
  - Remove stop words
  - Lemmatize text
  - TF-IDF Vectorization
- Unweighted Model:
  - Fit the data according to the percent of instances of each band
- Weighted Model:
  - Specify prior probabilities to improve accuracy for less common bands

Unweighted Results	Weighted Results
73.55%	69.70%



05 Combined Method

- Combine HDBSCAN and Band Classification (weighted) to filter HDBSCAN's predictions with Band Classification predictions
- Yields more precise results, predicting fewer and narrower "areas of interest"

Combined Method	
Predicts >= 1 "Area of Interest" for Projects	Measurements Captured per Project
67.17%	44.72%

Project Code	***HDBSCAN* Prediction**	***Band Classification* (*weighted*) Prediction**	***Combined (unweighted)* Prediction**
2017.1.00786.S	Band: [Frequency Range] 3: [89.105 101.005] 3: [109.775 115.160] 6: [213.095 220.395] 6: [227.095 231.490] 7: [355.090 357.225] 7: [344.980 345.180] 7: [345.785 345.815]	Bands: 6 7	Band: [Frequency Range] 6: [213.095 220.395] 6: [227.095 231.490] 7: [355.090 357.225] 7: [344.980 345.180] 7: [345.785 345.815]

Limitations & Assumptions

- Limitations
  - Difficulty measuring success
  - Did not have full research papers to train on
- Assumption
  - All “areas of interest” already exist in the data
  - LDA topics are salient and discriminant
  - Optimal weights were calculated

Conclusions

- Combined (weighted) process is useful
  - Researchers submitting projects to ALMA
  - Proposal reviewers
  - Understanding of project proposals
  - Recommendations for proposed projects
- Applicable and adaptable to telescopes beyond ALMA
  - Expand Line/Continuum classification to Very Large Array (VLA) in New Mexico

Acknowledgements

Antonios Mamalakis, Faculty Mentor

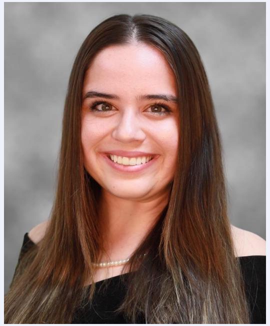
Adele Plunkett, Sponsor

Capstone Team

Arnav Boppudi

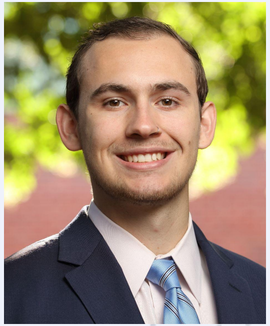
Ryan Lipps

Kaleigh O'Hara



Noah McIntire

Brendan Puglisi



Citations

LDA Graphic:

M. Bakrey, "All About Latent Dirichlet Allocation (LDA) in NLP," Medium, 01-Nov-2020. [Online]. Available: https://mohamedbakrey094.medium.com/all-about-latent-dirichlet-allocation-lda-in-nlp-6cfa7825034e. [Accessed: 26-April-2024].

HDBSCAN Example Graph Code:

"HDBSCAN clustering with sklearn," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_hdbscan.html#sphx-glr-auto-examples-cluster-plot-hdbscan-py. [Accessed: 26-April-2024].