

# Improvements on NRAO Proposal Classification Model

Riya Pulla, Krishna Kumar, Carter Day

School of Data Science, University of Virginia, Charlottesville, VA, USA

wbd3kw@virginia.edu, dcy2dq@virginia.edu, yuk7du@virginia.edu

5/2/2025

## Abstract

In cooperation with the National Radio Astronomy Observatory (NRAO), this project builds upon previous work in developing a machine learning pipeline to streamline proposal submissions for the Atacama Large Millimeter Array (ALMA). This year, we aim to enhance the accuracy and usability of frequency range recommendations provided to astronomers and create a user interface for scientists to use. Analytically, the team focuses on optimizing the topic modelling segment of the existing pipeline by incorporating astroBERT, an astrophysics semantic language model, and comparing various clustering techniques to find which model provided the best average measurement hit rate per project. We find that performance does not necessarily show improvement when incorporating astroBERT - the best combination features Latent Dirichlet Allocation and Spectral Clustering, with an average hit rate of 96 percent. The team also successfully developed an interactive exploratory dashboard allowing researchers to visualize key aspects of their proposal such as topic clusters, common frequencies, and associated chemical frequencies.

## 1 Introduction

### 1.1 Background

The ALMA Observatory is a radio telescope array located in the Atacama Desert in Chile.

ALMA is the largest astronomical project in existence and is open to anyone in the world to use via a project proposal process. Each submitted project proposal requires a technical plan outlining the specific settings and measurements the research team wishes to make. The design of the telescope array allows ALMA to detect electromagnetic radiation on a fairly continuous range of frequencies from 35 GHz to 950 GHz at two-decimal-point precision, divided into ten discrete bands. Furthermore, the technical specifications of the telescope allow individual measurements to span a range of approximately 4 GHz. As ALMA is limited to accepting 400 proposals throughout each year, access to the telescope is both highly sought-after and competitive. As such, designing clear, technically proficient proposals is paramount for research to optimize their chance to gain access to such resources.

### 1.2 Summary of Previous Project

In 2024, NRAO collaborated with University of Virginia's School of Data Science to develop a machine learning pipeline that provided targeted frequency/band recommendations for researchers to specify in their proposal submission to ALMA based on their project's stated goals and technical requirements. The student capstone team developed a multi-step model that first classifies proposals as either "line" or "continuum" types using logistic regression, with "line" proposals being the focus of further recommendations due to their specificity. Next, accepted line proposals are grouped into topics

---

\*Capstone Final Paper.

using Latent Dirichlet Allocation (LDA), with measurements within each topic then clustered using HDBSCAN to identify common frequency usage patterns. Simultaneously, a Multinomial Naive Bayes classifier predicts the most relevant ALMA frequency bands for each proposal based on the abstract and title. These predictions are then combined with the HDBSCAN clusters to recommend targeted “areas of interest”—narrow frequency ranges likely to be valuable for the proposed research. They also engineered features like TF-IDF vectorized text for the proposals and used topic-word distributions to improve interpretability. Performance evaluation shows that the final combined approach accurately predicts at least one measurement for 67% of the test projects with an average hit rate of 45%, while individually the LDA + HDBSCAN approach yielded an average hit rate of 88%.

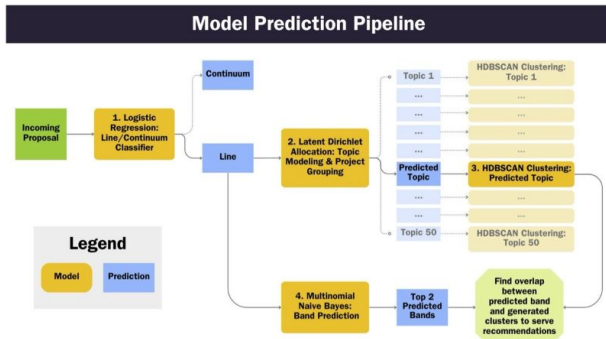


Figure 1: The machine learning pipeline developed by the 2023-2024 UVA MSDS capstone team.

Upon conclusion of their work, the previous group noted that there may be merit in investigating future optimizations to their existing modeling approaches. One such suggestion included the incorporation of astroBERT, a semantic language model fine-tuned to parse NASA projects in their Astrophysics Data System (ADS). This year, the new capstone team acts on this suggestion, examining the performance impact of astroBERT-generated embeddings on topic modeling and clustering approaches to proposal classification. The group also looks to enhance the dashboard created by

the previous team, adding relevant chemical information such as observed frequencies and activation energies.

## 2 Data Discussion

Our data consisted of accepted ALMA project proposals, filtered to exclude those with continuum setups or more than 2 band overlaps. Measurements were parsed from frequency ranges into median values to simplify modeling. Although these data provide reliable technical details, they lack the full textual richness of full proposals.

### 2.1 Full Text Proposals and Data Security

We were unable to access the full text of the proposal due to privacy concerns, particularly about training models on proprietary research proposals. A potential future direction involves creating a secure, access-controlled environment where full proposals can be used with researcher consent.

### 2.2 NLP with AstroBERT

AstroBERT is a transformer-based NLP model pre-trained in astrophysical papers from the Astrophysics Data System (ADS) [1]. Unlike LDA, which uses word co-occurrence, AstroBERT captures semantic similarity between proposals at a sentence and document level [3]. We used the final-layer CLS token embeddings from AstroBERT as vector representations of each proposal. Although AstroBERT did not outperform LDA in all clustering settings, it provided interpretable embeddings that can be used to explore further classification or fine-tuning in the future.

### 2.3 Text Pre-processing and Vectorization

For LDA, we followed the same preprocessing steps as the previous capstone team: lowercasing text, removing punctuation, filtering stopwords, and lemmatizing words to their base forms.

These steps are essential for reducing noise and ensuring consistency in traditional bag-of-words models. We also experimented with removing domain-specific terms like “galaxy”, “star”, and “observation”, in case they might dilute topic quality due to their high frequency. However, this had little effect on model performance and occasionally reduced interpretability.

## 3 Methodology

### 3.1 Model Updates

In our project, we built upon the previous years capstone’s multi-step model pipeline by experimenting with alternative topic modeling and clustering techniques. The original model used Latent Dirichlet Allocation for topic modeling, HDBSCAN for clustering frequency measurements, and a Multinomial Naive Bayes classifier for frequency band prediction. Our updates focused on replacing and improving the accuracy of the topic modeling competent. We also wanted to see how a language model trained on astrophysics text, like AstroBERT, could improve our results.

#### 3.1.1 Topic Modeling

The original project used LDA to assign each ALMA proposal to one of 50 topics based on the proposal title and abstract. We evaluated this method against AstroBERT, a domain-specific transformer-based language model pre-trained on astrophysics literature. AstroBERT generates dense semantic embeddings for each proposal, capturing deeper contextual meaning than LDA’s word-count-based approach. These embeddings were then used as input for downstream clustering. Our hypothesis was that AstroBERT would result in a more coherent and semantically meaningful topic groupings.

#### 3.1.2 Clustering

We tested a variety of clustering algorithms to pair with both LDA and AstroBERT outputs. We tested HDBSCAN as well for consistency

and to compare performance of additional models. We evaluated OPTICS, Spectral Clustering, KMeans, Gaussian Mixture Models, and Dirichlet Process Clustering.

#### 3.1.3 Additional Data

We used 3,178 filtered ALMA proposals and 23,482 measurements to train our model. We introduced new engineered features such as chemical species mentioned in submitted proposals, median frequency approximations, and normalized measurement counts. These features allowed us to evaluate topic cohesion and measurement overlap across clusters, adding a layer of interpretability to our results.

#### 3.1.4 Improvement Criteria and Metrics

The primary metric for evaluating model quality was the average measurement hit rate per project; the proportion of a project’s true measurement frequencies that fall within its predicted areas of interest. We also considered precision of frequency band prediction, topic coherence (qualitatively), and the percentage of proposals with at least one matched measurement.

#### 3.1.5 Additional Methods

A few clustering methods tested here do not support traditional inference on test data based on their natural limitations:

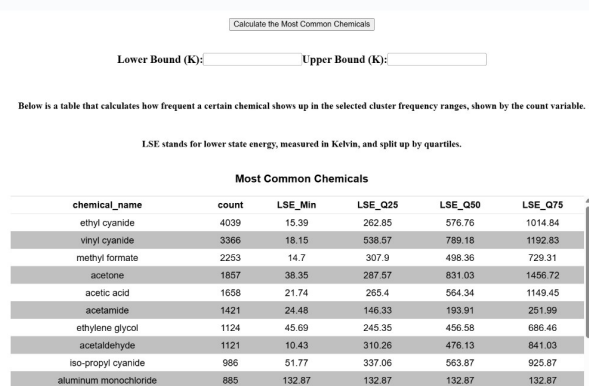
- **HDBSCAN** is a density-based algorithm and adding predicted points can upset the existing density balance among generated clusters.
- **Spectral Clustering** relies on Eigenvalue decomposition and therefore would require recomputation for every additional point.

These limitations are relevant for testing astroBERT’s topic modeling performance as the the project data includes a train-test split. Multiple proxies for inference were considered such as training a neural network on the generated clusters and training classifications. For the sake of time and simplicity, the team opted to assign the

test project data to clusters based on Euclidean distance to nearest centroid of a given cluster.

### 3.2 Updating Dashboard Information

The current dashboard gives the user the ability to select a topic from the topic generation part of the model, as well as one or up to ten of the ALMA frequency bands. This displays all of the project’s clusters for that particular topic and bands, as well as displaying a table of a selected cluster within that graph. This table shows all the different projects in that cluster, as well as their frequency ranges and some miscellaneous project information. As a part of the deliverables for this project, an addition to the current dashboard for the project was required. A new table was created which displays the chemicals that were most likely found in a certain range of frequencies. This can be calculated for any selected cluster on the dashboard. It can also be bounded by the lower state energy of the chemical in Kelvin, and the quartile ranges of the different chemicals’ lower state energy are shown in the table. The goal for this dashboard is to make it easier for the astronomers and researchers using this tool and to help them make better decisions when it comes to setting up their project. If they are able to see a list of chemicals like this, it may give them new ideas on what to look for or what they should expect if they point the telescopes in that frequency range.



Calculate the Most Common Chemicals

Lower Bound (K):  Upper Bound (K):

Below is a table that calculates how frequent a certain chemical shows up in the selected cluster frequency ranges, shown by the count variable.

LSE stands for lower state energy, measured in Kelvin, and split up by quartiles.

Most Common Chemicals					
chemical_name	count	LSE_Min	LSE_Q25	LSE_Q50	LSE_Q75
ethyl cyanide	4039	15.39	262.85	576.76	1014.84
vinyl cyanide	3366	18.15	538.57	789.18	1192.83
methyl formate	2253	14.7	307.9	498.36	729.31
acetone	1857	38.35	287.57	831.03	1456.72
acetic acid	1658	21.74	265.4	564.34	1149.45
acetamide	1421	24.48	146.33	193.91	251.99
ethylene glycol	1124	45.69	245.35	456.58	686.46
acetaldehyde	1121	10.43	310.26	476.13	841.03
iso-propyl cyanide	986	51.77	337.06	563.87	925.87
aluminum monochloride	885	132.87	132.87	132.87	132.87

Figure 2: The most common chemicals table from the dashboard with example data.

## 4 Results

### 4.1 LDA Results

Below is the hit rate comparisons for the individual clustering methods tested while maintaining the original project’s Latent Dirichlet Allocation Model. The first row represents the result of the original work.

Topic Generation	Clustering	Avg. Hit Rate
Latent Dirichlet Allocation	HDBSCAN	88%
Latent Dirichlet Allocation	OPTIC	86%
Latent Dirichlet Allocation	Spectral Clustering	96%

From the above results, it is clear the Spectral Clustering significantly outperforms the tested alternatives. Given this, the team elected to use Spectral Clustering as the method of choice for clustering within the individual topics when testing the performance of astroBERT.

### 4.2 astroBERT Results

Below are the hit rate comparisons for various topic modeling approaches using astroBERT-generated embeddings.

Topic Generation	Clustering	Avg. Hit Rate
astroBERT+ <i>HDBSCAN</i>	Spectral Clustering	49%
astroBERT+ <i>Spectral Clustering</i>	Spectral Clustering	43%
astroBERT+ <i>K-Means</i>	Spectral Clustering	86%
astroBERT+ <i>Gaussian Mixture Model</i>	Spectral Clustering	87%
astroBERT+ <i>Dirichlet Process</i>	Spectral Clustering	88%

Generating topics with natural inference methods proved much more effective than methods that required prediction proxies. However, we find in general that the use of astroBERT embeddings with clustered topic generation does not significantly improve the model’s performance when compared to the original methodology.

### 4.3 Discussion

Overall, the team recommends that the model retains Latent Dirichlet Allocation as its method for topic generation while updating the clustering within each topic to use Spectral Clustering over HDBSCAN for improvements on measurement hit rates. That said, there remains some potential analytic value in using some of the astroBERT methodologies described above. In particular, the probabilistic assignments of Gaussian Mixture Models and Dirichlet Processes allow for projects to be categorized under multiple topics, allowing for more robust and salient explainability while also providing practical value, as in reality, many projects do not often fit into one exact category.

## 5 Future Work

### 5.1 Full Text Proposals

As mentioned above, we were not able to procure the full text proposals due to data privacy concerns, but if that were to change any time in the future with an access-controlled environment where researchers can safely share their proposals, including them in the training of the models we believe would boost performance. Unfortunately, in the current climate and the foreseeable future, we don’t see that changing, and working with the abstracts and frequency ranges is the best we can do in terms of data fed into the pipeline.

### 5.2 Modeling Techniques

We believe that the model pipeline as it currently exists maximizes the potential predictive power

of the data we are given. We tested many different topic generation and clustering techniques, and we believe there is little room for improvement there in the way the model is currently set up. If optimization past what this model pipeline gives is needed, a look into large language models and how they use clustering and predictive power could be insightful. If there is any improvement to come from large language models, we believe it would only be a marginal gain, but as time goes on and large language models are improved, it might be a topic to explore.

### 5.3 Improved Dashboard

The last part that could be improved in the project is on the front end with the dashboard. While it currently presents data effectively, there is potential for it to be better by making it more interactive and insightful. An idea is to go deeper into the analysis to provide more precise recommendations on specific molecules based on the cluster data and previous projects. This could involve flagging chemicals that exhibit promising characteristics or unusual activity and highlighting those that might warrant further investigation. Anything that might make it easier for the astronomers to make well informed decisions and potentially provide them with new ideas is something to strive for.

## References

- [1] Lhoest, Quentin, Albert Villanova del Moral, Yacine Jernite, Nicolas Patil, Patrick von Platen, Clara Ma, Mariama Drame, Julien Plu, Lewis Tunstall, Teven Le Scao, Victor Sanh, Canwen Xu, Sylvain Gugger, Lena Marvin, and Thomas Wolf. *Datasets: A Community Library for Natural Language Processing*. arXiv preprint arXiv:2112.00590, 2021. <https://doi.org/10.48550/arXiv.2112.00590>
- [2] Hankar, M., Kasri, M., & Beni-Hssane, A. (2025). *A comprehensive overview of topic modeling: Techniques, applications and challenges*. Neurocomputing,

628, 129638. <https://doi.org/10.1016/j.neucom.2025.129638>

- [3] Arnav Boppudi, Ryan Lipps, Noah McIntire, Kaleigh O'Hara; Brendan Puglisi, Antonios Mamalakis *Optimizing the ALMA Research Proposal Process with Machine Learning*. 2024 Systems and Information Engineering Design Symposium (SIEDS) <https://doi.org/10.1109/SIEDS61124.2024.10534693>