NRAO Capstone Checkpoint 2

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

2023-4-24

Stakeholders

- Adele Plunkett, Sponsor
- Antonios Mamalakis, Mentor
- Arnav Boppudi, Team member
- Brendan Puglisi, Team member
- Kaleigh O'Hara, Team member
- Noah McIntire, Team member
- Ryan Lipps, Team member

Abstract

Every year, astronomers from around the world submit research proposals to the Atacama Large Millimeter Array (ALMA), the largest radio telescope array in the world. The aim of the current work is to streamline the proposal process for astronomers submitting projects to ALMA by suggesting frequency ranges that may be relevant to their research based on their proposal text. We introduce a pipeline of supervised and unsupervised machine learning models, each using various representations of the title and abstract of an incoming proposal. First, a logistic regression filters out proposed projects that are not expected to need specific technical setups. Second, if a technical setup is deemed necessary, our pipeline assigns an incoming project to one of 50 "similar project" groups, defined by topics generated from Latent Dirichlet Allocation (LDA). Third, we apply Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to mine patterns in measurements ("areas of interest") made in previous projects, for each one of the 50 "similar project" groups. In parallel to the aforementioned topic modeling and HDBSCAN mining, we employ a Multinomial Naive Bayes classifier to predict the broad frequency range defined by the technical limitations of ALMA (frequency band) that we expect a project to make measurements in. Finally, we offer researchers a list of the mined "areas of interest" filtered by the predictions of the Multinomial Naive Bayes classifier. Ultimately, given a proposed project title and abstract, our pipeline generates several recommended "areas of interest" that one should consider measuring in.

Regarding the performance of our models, we find that 67.17% of test projects match at least one of the recommended "areas of interest", with an average hit rate of 44.72% across measurements within each test project, when limiting to the top two band predictions. When we disregard band predictions, 88.81% of test projects match at least one recommended "area of interest" with an average hit rate of 60.00% across measurements within each test project. While the latter approach gives better results in terms of hit rates, we believe the combination of models provides a good balance of accuracy and precision.

Project Outline

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. This can be daunting since 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. This continuous range of frequencies is broken up into ten discrete bands. Furthermore, the technical specifications of the telescope allow individual measurements to span a range of approximately 4 GHz. On top of the technical nature of project proposals, ALMA can only accept around 400 proposals to enact throughout the year, making it a valuable and competitive resource. Thus, it is paramount that researchers submit well designed and technically proficient proposals to increase their chances of acceptance.

Our work streamlines the ALMA proposal process by providing researchers insight and assistance on how to set their measurement requirements in the project proposal technical plan, with the intent of optimizing the insight they can extract from the data.

Our process involves two subtasks — one to classify the type of project based on expected technical plan, the second to provide recommendations for a possible frequency to observe, based on information from task one.

The first subtask in our work classifies proposed projects as either "line" or "continuum" setups. For ALMA's technical specifications, "line" setups focus on detecting specific frequencies associated with spectral lines of molecules, revealing chemical compositions and physical conditions of celestial objects. "Continuum" setups measure broad-spectrum emissions to understand the energy output, temperature, and structure of astronomical bodies. Because continuum setups are often used for specific projects and can cover a wide range of frequencies in a single observation, we work solely with line setups when providing recommendations.

The second, more extensive task of this approach, involves creating an interpretable suite of models that assist in recommending frequency ranges for line setups, based on a project's title and abstract. This provides stakeholders with a tool to explore similar accepted projects based on the perceived topic of their proposal and the related frequencies that may be useful to observe. The full model prediction pipeline is shown in Fig.

Outline of the Data

The data has been collected over the course of 10 years, from the use of the ALMA observatory. It has been sourced from previous studies utilizing the antennas. This includes the abstract of the experiments along with the positioning of the antennas (e.g., right ascension, declination, and band focus). If needed, full proposal data can be accessed, however this data is not publicly available until a year after the proposal is accepted.

Target variables:

- Document feature extraction based on titles and abstracts for proposal clustering
- Array frequency ranges
  - Each observation in the data includes the frequency ranges used to observe spectral signatures of objects in space
  - Each observation for the array must fall within certain observable frequency bands and specific frequency ranges within those bands relate to the observable signatures of molecules
  - We will attempt to recommend the optimal frequency band and range based on similar research proposals, maximizing the importance and/or number of interesting frequency ranges and molecules to observe

Success Criteria

Success Criteria	Description
SC1	An algorithm that recommends the optimal technical setup based on the text that an astronomer specifies in the proposal, specifically the title and abstract. The primary goal will be to optimize the frequency ranges to be observed simultaneously in a project
SC2	Github Repo housing code
SC3	Algorithm and supporting code organized into python package
SC4	Reach goal for algorithm: suggest alternative data to collect within recommended frequency band and array settings

Assumptions and Limitations

Identifier	Description
Limitation	There is no set or efficient way to measure the accuracy or success of our model, unless we make a train/test split on the data. Since there is not a lot of data we need to be conscientious about the size of the split to optimize training information and testing capability.
Limitation	Title and abstract may not be sufficient for model training, this may require full proposal text for better accuracy, however, proposal text is not publicly available
Assumption	The project data can be "labeled" in order to cluster/develop a model to suggest frequency setups
Assumption	Additional metadata, e.g. keyword, can be used to assist model performance, keyword data only available at proposal submission. This is to say we will not have these keywords immediately, but maybe we can ask users for them during proposal submission
Assumption	There is not one single "correct" tuning per proposal — multiple tunings could work depending on researchers' preferences

Potential Background Literature and Resources

- ALMA Science Archive: https://almascience.nrao.edu/aq/
- ALMA Proposal Preparation (start at Slide 36 for spectral setup): https://science.nrao.edu/facilities/alma/naasc-workshops/nrao-cd-antofagasta23/presentations-1/07-OT_Cycle10-AmbassadorsProgrampptx.pdf
- ALMA Proposer's Guide: Spectral capabilities: https://almascience.nrao.edu/proposing/proposers-guide#autotoc-item-autotoc-69
- ALMA Science Archive Users Manual: https://almascience.nrao.edu/documents-and-tools/cycle9/science-archive-manual
- ALMiner Tool (for programmatic access of ALMA data/metadata): https://www.alma-allegro.nl/alminer/
- End-to-End Multiclass Text Classification (NLP): https://www.analyticsvidhya.com/blog/2021/11/a-guide-to-building-an-end-to-end-multiclass-text-classification-model/

Initial Model Steps:

- A language model which extracts features from the research proposal
- A clustering model to group similar features from the language model and recommend optimal or frequently used "settings" for accurate positioning of antennas and to optimize frequency support

Modeling and Analysis Checkpoint

To begin the modeling process we focused on building a binary classifier based on the project texts and abstracts. Each member built a different model and we picked the one with the highest accuracy. The models considered were as follows:

- Logistic regression using TF-IDF vectorization of text
- Shallow neural network using TF-IDF vectorization of text
- COP-KMeans using TF-IDF vectorization of text
- Latent Dirichlet Allocation using count vectorization of text
- Support Vector Machine using count vectorization
- Random Forest using count vectorization

The best model from this step was the logistic regression classification model based on a TF-IDF vectorization of the project titles and abstracts. We used this model to distinguish between line and continuum projects. The most important metric for this model was accuracy (0.89, but varies by ~.02 based on test train split) with additional tuning based on the confusion matrix. Based on the interpretability and accuracy of this model, we feel comfortable delivering it to stakeholders at NRAO and other astronomy organizations to reduce resources allocated to ensuring projects are correctly classified.

We employ a parallel modeling approach to recommend the optimal frequency setup for incoming projects: one path using Latent Dirichlet Allocation to group similar projects by topics, which we then use to cluster measurements using HDBSCAN, and one path using Multinomial Naive Bayes to predict the top two most likely frequency bands for a project.

The LDA/HDBSCAN and *Band Classification (weighted)* methods adequately predict "areas of interest" for each project in the test set. HDBSCAN provides many diverse, typically narrow, areas of interest while *Band Classification (weighted)* provides two wide areas of interest (bands). In an attempt to maintain the narrowness of HDBSCAN's predictions while benefiting from the reduced number of "areas of interest" in *Band Classification (weighted)*'s results, we combined these two approaches. We refer to this approach as the *Combined (weighted)* approach. In this approach, we find the intersecting areas of interest in the topic predicted by HDBSCAN and the *Band Classification (weighted)* for each project. See an example in Table **??**.

This approach is more precise than the HDBSCAN approach since it predicts fewer frequency ranges for each project. Furthermore, this approach is also more precise than the *Band Classification (weighted)* approaches since it predicts narrower area of interests.

We found that across the testing data, the *Combined (weighted)* approach predicts at least one correct measurement for 67.17% of the test set projects with an average of 44.72% of the measurements in a test project captured in the prediction.

Raw Data:

- Abstract and titles of studies, keywords, positioning of antennas (right ascension, declination, band, continuum sensitivity, frequency support, etc.)
- All the data is easily accessible via https://almascience.nrao.edu/aq/, and can be queried into a jupyter notebook or python script using the ALminer package https://alminer.readthedocs.io/en/latest/.

Target variables:

- Feature extraction from proposal language processing
- Frequency ranges for similar research proposals
- Predicted Bands
- Engineered median frequency of measurements

Potential predictor variables:

- Project title and abstract (possibly full project text)

- Keywords related to proposal

Possible challenges:

- Somewhat small sample of project abstracts/ titles, limiting train/test opportunities (roughly 4000)

Project Timeline

Project Stage	Due Date
Presentation/tutorial on data and use from Dr. Plunkett	11/17/23
Literature review	1/17/24
Initial NLP pipeline (cleaning, tokenizing, lemmaization, etc.)	2/7/24
Initial NLP model	2/7/24
Initial classification model	2/7/24
SIEDS proposal	2/7/24
Model improvements and other architecture exploration	3/31/24
SIEDS paper due	4/7/24
Prototype for next ALMA research proposal deadline	3/15/24 - 4/15/24
Comprehensive Testing/ Evaluation of model(s)	3/15/24 - 4/15/24
SIEDS Conference	5/3/24
Submit Python Packaging and Github Repo	4/30/24

Potential Concerns and Blockers

Identifier	Description
Concern	No immediate experience or knowledge of NLP
Concern	Technical knowledge of frequency setups, general knowledge of data and proposals
Blocker	Archives go down periodically. There are three mirrors of the archives, so we may have to change which server we are sourcing data from. This amounts to changing the URL used in ALminer or other data accessing platform. Max expected downtime is 1 hour.

Future Work Plan

We will package our code into a notebook or python script that hosts a simple but effective dashboard that deploys our model and provides researchers with visualizations of the recommended setups.

Reflection Questions

Reflection question 1: What was the biggest challenge that you faced with this project?

The biggest challenge of this project was understanding the nuances of both the overall problem and the data. The nature of the problem is vast and undefined, as the projects on which we must base predictions have no specified targets. Ultimately, the project requires recommending "optimal" settings where we have no insight on why previous projects were set up the way they were.

Reflection question 2 Did this project stretch you to grow? If so, how?

This project stretched us to grow in multiple ways. We grew into better teammates as we had to work as a team. We grew as data scientists as we had to come up with our own solution as opposed to implementing a single well-known solution. We had to research and learn a broad set of models beyond the scope of our coursework to construct a unique solution.

Reflection question 3 Do you believe the capstone experience will be helpful to your career? If so, how?

This experience was helpful to our career as we had to adapt to a problem that does not have a single well-defined solution. This involved making deliberate choices on data cleaning, processing, model selection, model tuning, evaluation, and presentation that did not have a "right answer". We also gained valuable experience in working for a client.

Reflection question 4 Anything else that you would like to share?

Overall, the bulk of capstone should start earlier. There is not enough time to do the lion's share of the work in the spring. This means either starting capstone in the summer and/or having more strict data access, cleaning, etc. in the fall.

If you would like groups to do SIEDS in the future, more guidance and resources on academic writing and poster-making would be valuable.