NRAO Capstone Project Proposal

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

2023-11-28

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

The ALMA Observatory in Chile is the largest astronomical project in existence, and is open to anyone in the world to use. The aim of this project is to streamline the proposal process for astronomers who use ALMA. The Capstone team will develop an algorithm that suggests the optimal technical setup based on the text that an astronomer specifies in their proposal. More than 4000 projects have been undertaken in the first 10 years of ALMA operations, and all projects are visible (with all metadata, and about 90% of the scientific data available) via the user interface of the ALMA Science Archive: https://almascience.nrao.edu/aq/.

Project Outline

The aim of this project is to streamline the proposal process for astronomers who use ALMA. The Capstone team will develop an algorithm that suggests the optimal technical setup based on the text that an astronomer specifies in the proposal, specifically the title and abstract. The metadata for these projects can be used to train and test the model. The primary goal will be to optimize the frequency ranges to be observed simultaneously in a project. Additional technical setup requirements can be explored, if there is time and interest among the group. As astronomers can spend weeks preparing these proposals, and we suspect that the technical setup is not always optimized, important outcomes of this project include an immense time-savings for astronomers around the world, as well as a more efficient usage of the ALMA telescope for observations of the Universe. The stakeholders in this project include sponsor Adele Plunkett, the NRAO, ALMA Observatory, astronomers utilizing ALMA, and UVA capstone students.

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: add code into existing ALminer functionality
SC5	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 are continuing to explore other models focused on classifying projects to ultimately recommend the optimal frequency setup for incoming projects. So far the models we are exploring are:

- COP-KMeans - the semi-supervised nature of this model is potentially useful for tuning to classification accuracy
- Latent Dirichlet Allocation - allows for further mining of topics generated using Principal Component Analysis to generate "pure" topics
- Neural networks - allows for increased discretization and can classify one project into different categories, which may be useful for generalizing frequency tuning recommendations
  - Transformer models such as BERT
  - Non-transformer architectures

Much of the exploration surrounding these models is based on maximizing the precision of classification. This is due to the large number of frequency tunings used across projects, as the model is not useful for new proposals if it is suggesting too many measurements or too broad of a frequency range.

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

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	4/27 - 4/29
Submit Python Packaging and Github Repo	4/30/24

Future Work Plan

We will create a project classification model and then recommend combinations of median frequency and frequency difference based on this classification. The final frequency tuning recommendation will be done either using a neural network or using a grid to find all possible frequency tuning combinations within specified boundaries and sorting each project into one or many of these bins. We will create multiple models using various data cleaning methods. For example, we may remove combinations of median frequency and frequency difference from the list of classes that have only one 1 data point in the training set. Additionally, we may remove observations that are in the upper quartile of the number of measurements to reduce variability and to eliminate cases that appear to not add value to the model.

Upon completion of the initial model(s), we will evaluate and compare them. After selecting a model type, we will reassess data cleaning methods, hyperparameters, etc. to produce the best model.

Future Project Timeline

Project Stage	Due Date
Frequency setup recommendation model(s)	3/13/23 - 3/27/24
Comprehensive evaluation of recommendation model(s)	3/27/24 - 4/3/24
SIEDS paper due	4/7/24
Model improvements and other architecture exploration	3/15/24 - 4/15/24
SIEDS Conference	4/27 - 4/29
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.