Classification of Subreddits Using Comments

—

Done by the SEAmen
Darwis, Cui Cheng, Zul

Content

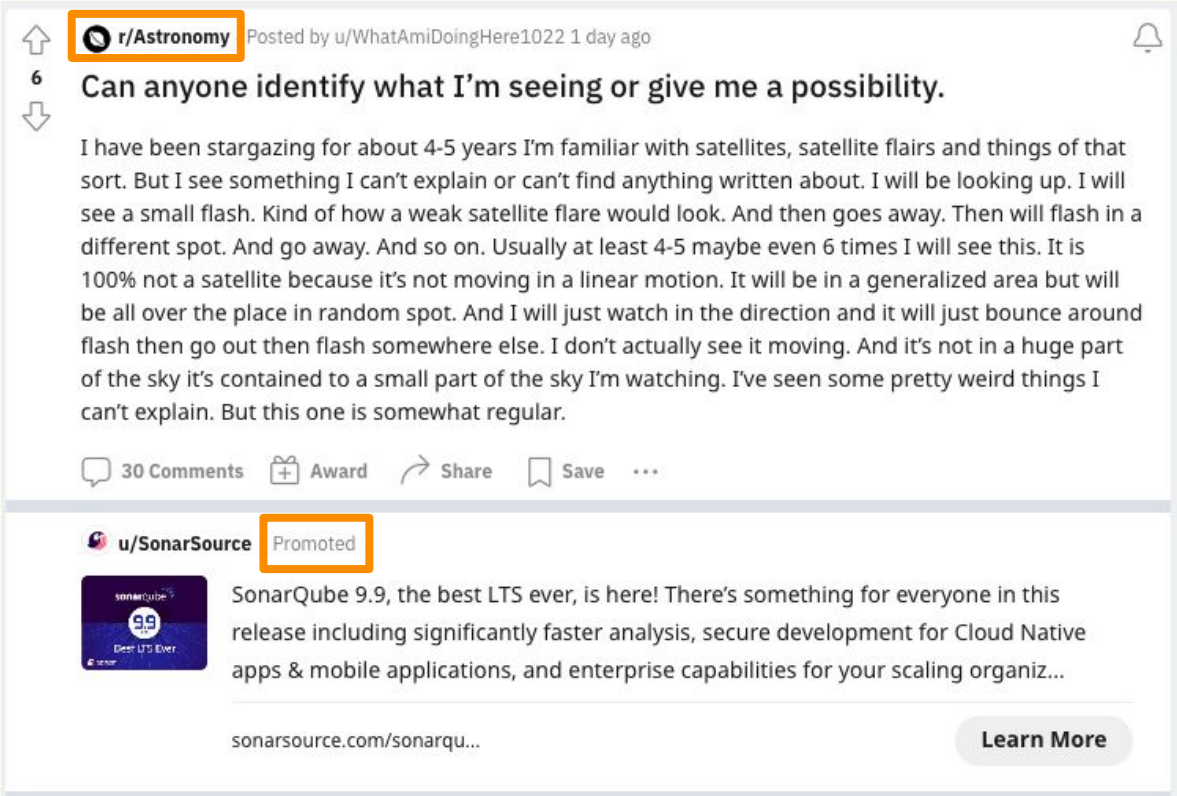
Background

HYPE FACTORY

- **Hype Factory**
- Influence Marketing Agency
- **Astronomers**
- Targeted advertising using Comments

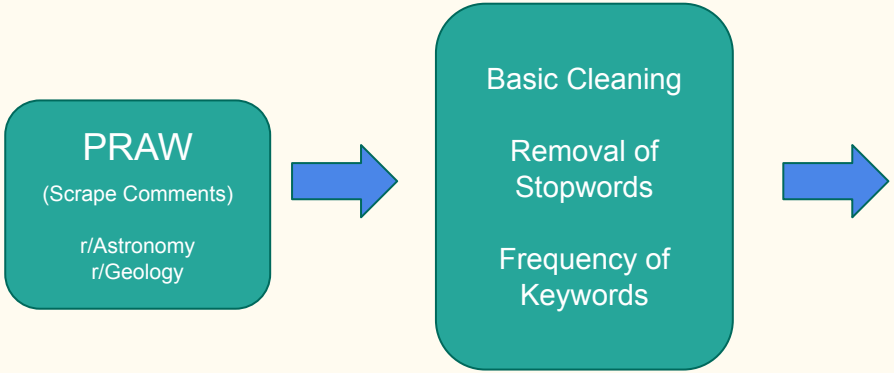
Background

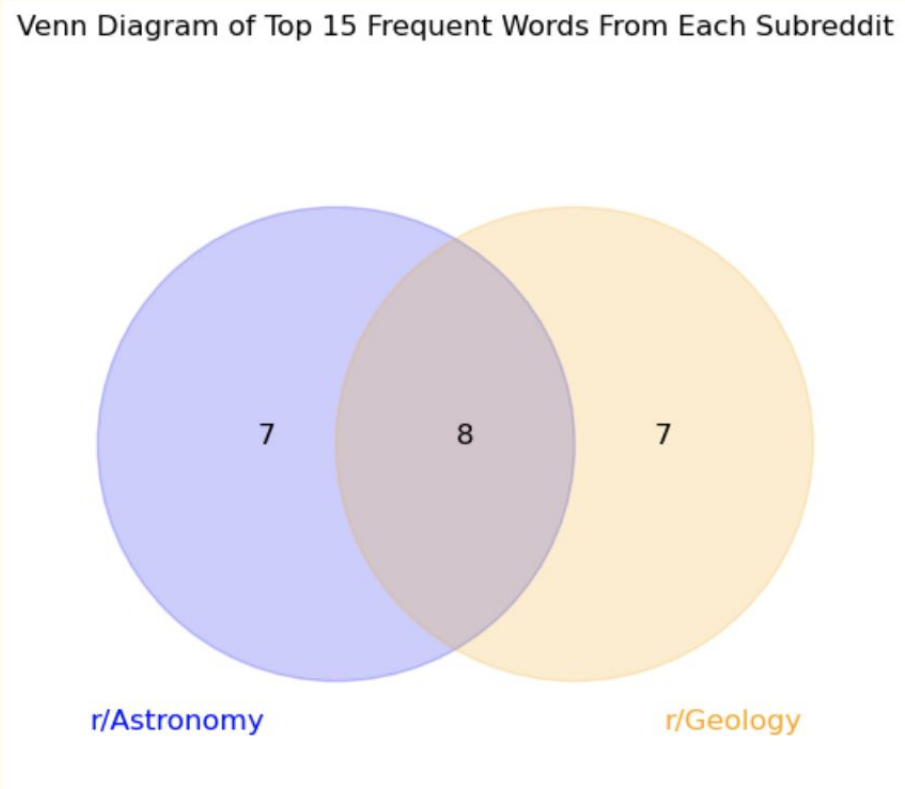
- r/Astronomy
- 2.8m members
- Significant TAM
- Comment Led Approach



Background

- Negative Feedback
- r/Geology (50%)
- Preliminary Investigations





Problem Statement

Build a Topic Classification Model to accurately classify comments between subreddits to better target the intended audience.

Stakeholders

- Managers at Hype Factory
- Our Clients

Workflow

Extraction

Data Extraction
(PRAW API)

Data Cleaning

Pre-processing

Tokenizing

Stemming
Lemmatization

Train - Test - Split

Word Vectorizing

Modeling

Iterative Modeling

Model Evaluation

Conclusions &
Recommendations

Data Cleaning

Hahaha, that was a very thorough explanation. I understand hwo telescopes work so muh more now. Thank youuuu! #astronomyexpert

Removing special fonts

unicodedata library

Data Cleaning

Hahaha, that was a very thorough explanation. I understand hwo telescopes work so muh more now. Thank youuuu! #astronomyexpert

Removing letters that are repeated more than 2 times

Removing repeated words

regex

Data Cleaning

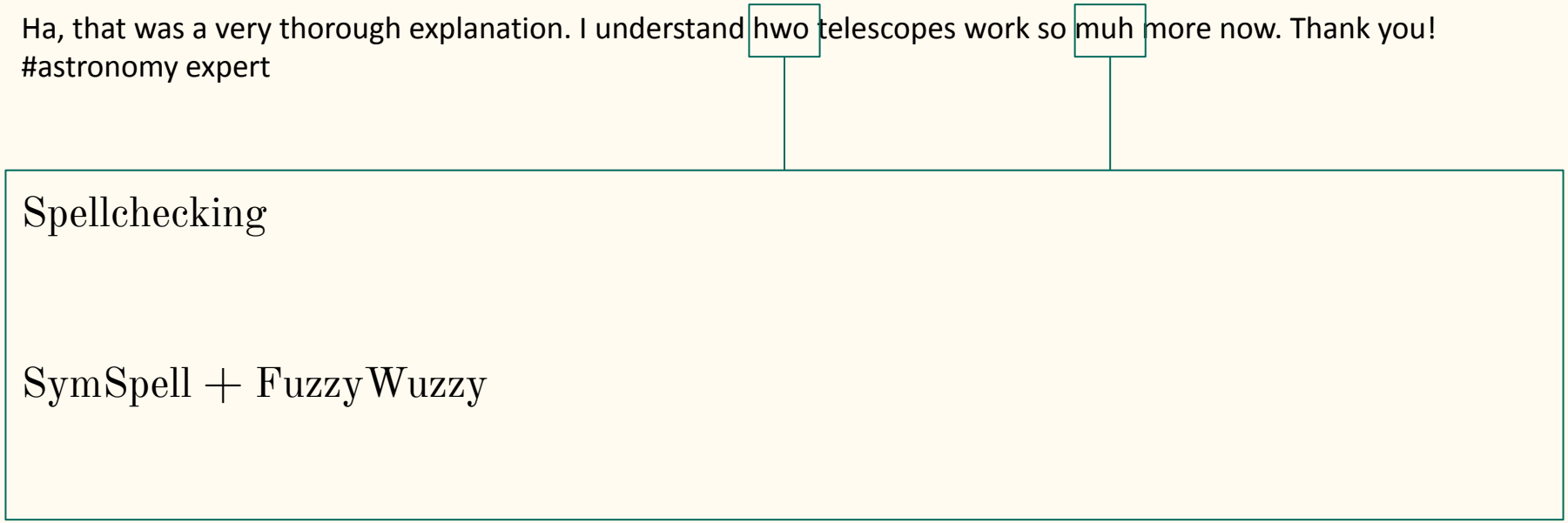
Ha, that was a very thorough explanation. I understand hwo telescopes work so muh more now. Thank you! #astronomyexpert

Splitting joined words

wordninja

Data Cleaning

Ha, that was a very thorough explanation. I understand hwo telescopes work so muh more now. Thank you! #astronomy expert

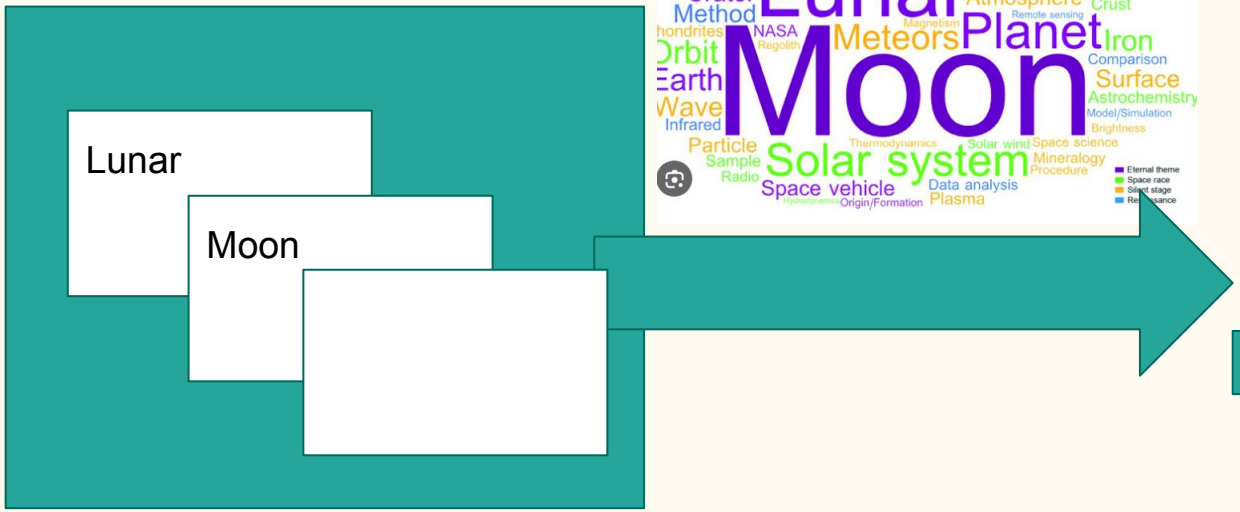


Data Cleaning

Ha, that was a very thorough explanation. I understand how telescopes work so much more now. Thank you! #astronomy expert

Hahaha, that was a very thorough explanation. I understand hwo telescopes work so muh more now. Thank youuuu! #astronomyexpert

Why is the algorithm underperforming?



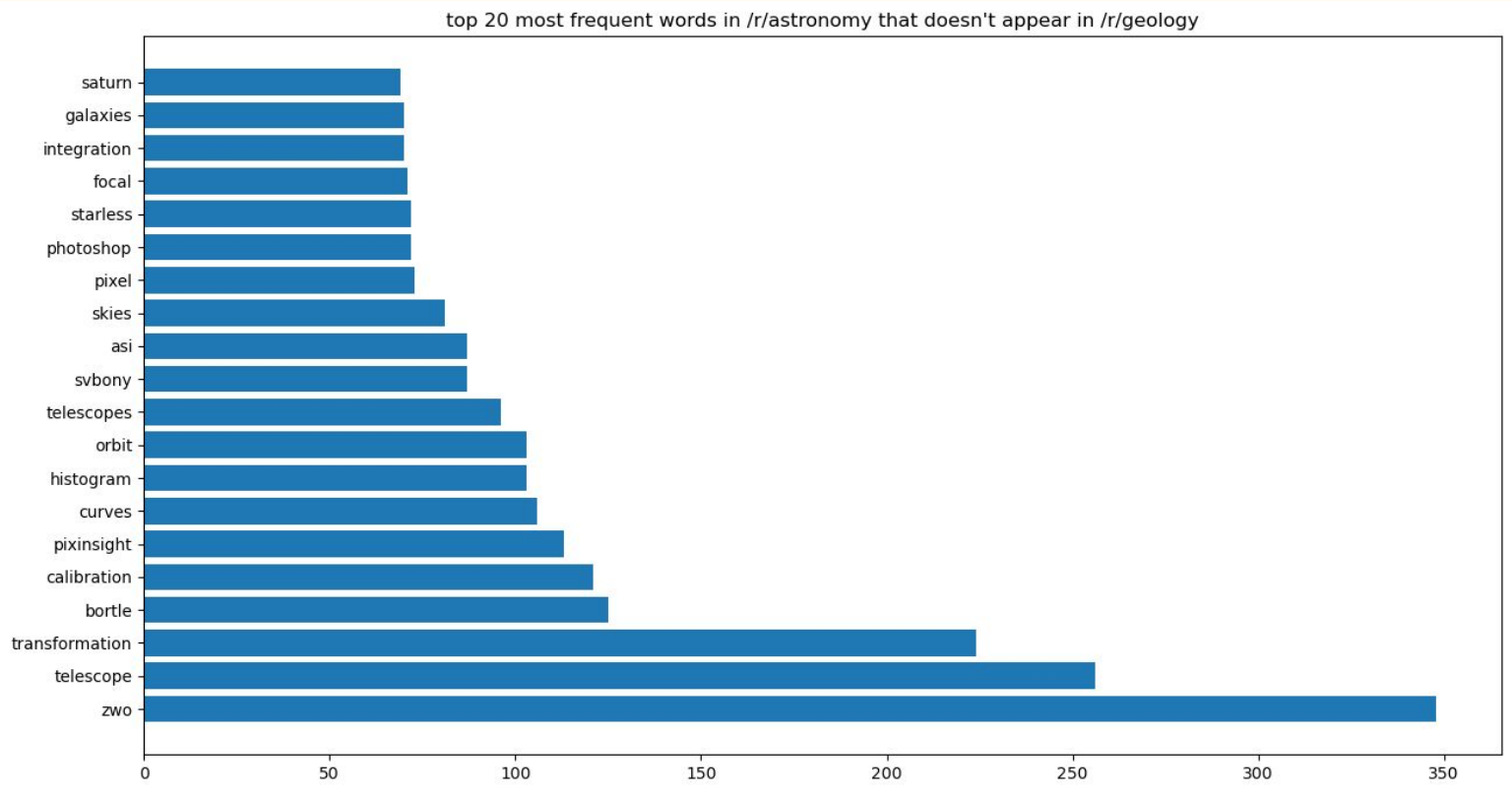
/r/Astronomy

Lunar

Moon

Others

Possible solution: Updating keywords

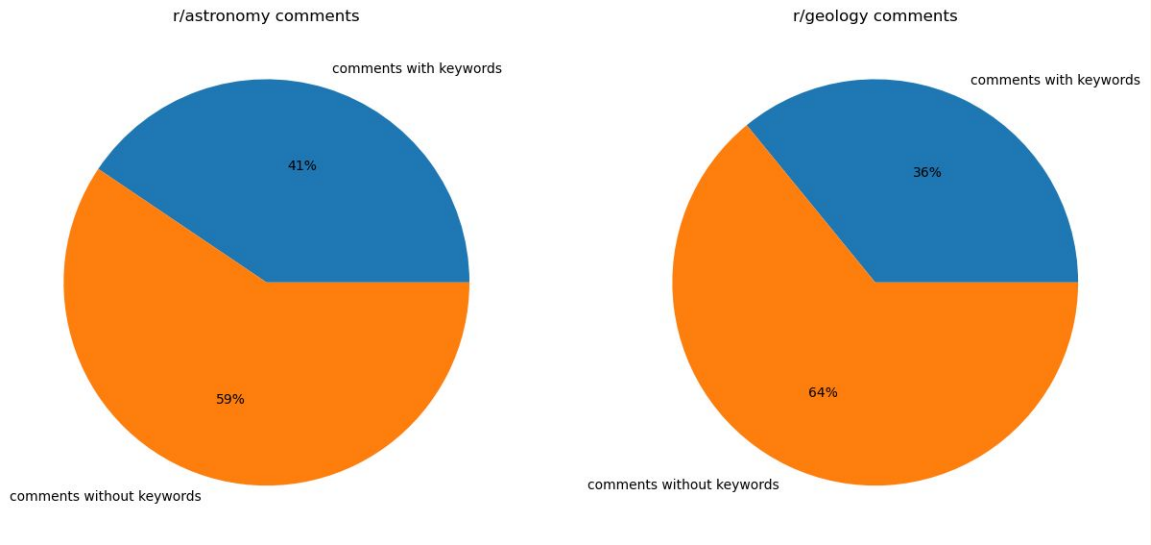


Why is the algorithm underperforming?

- Assumed that keywords distinct to each subreddit will be present in all the comments from each subreddit

Top 20 Keywords:

'zwo', 'telescope', 'transformation', 'bortle', 'calibration', 'pixinsight', 'curves', 'histogram', 'orbit', 'telescopes', 'svbony','asi', 'skies', 'pixel', 'photoshop', 'starless', 'focal', 'integration', 'galaxies', 'saturn'



Top 20 Keywords:

'geology', 'granite', 'clay', 'geologist', 'fossils', 'limestone', 'basalt', 'sedimentary', 'poll', 'sandstone', 'weathering', 'geologists', 'sediment', 'geologic', 'coal', 'calcite', 'deposits', 'metamorphic', 'igneous', 'pyrite'

Comments without keywords

['Well', 'u', 'would', 'get', 'alot', 'orange', 'ones']

['Go', 'wikipedia', 'several', 'references', 'linked', 'bottom', 'Berkshires', 'wiki']

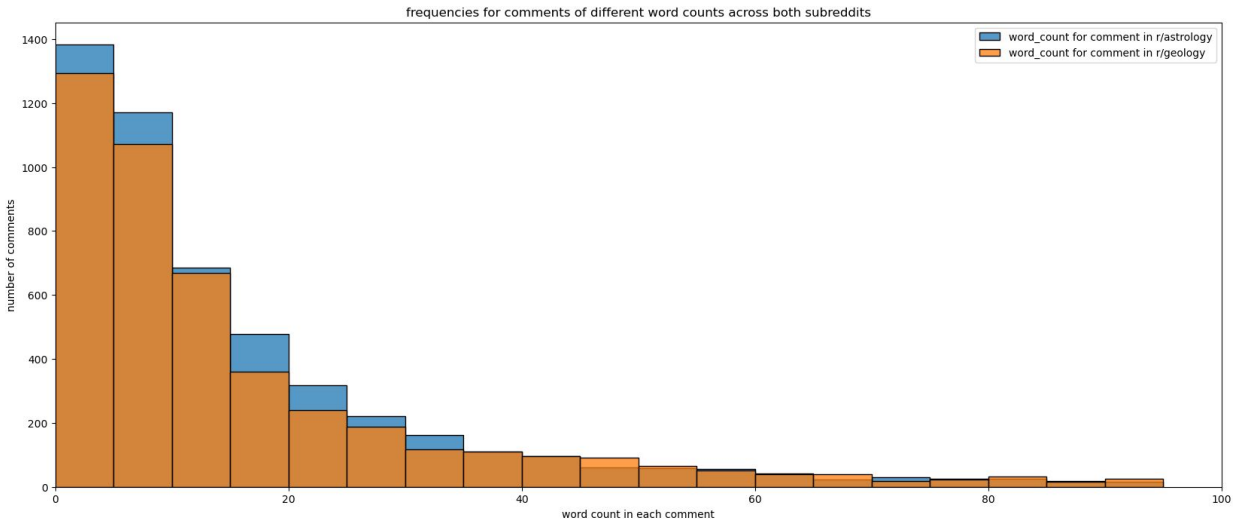
['one', 'done', 'volcanoes', 'form']

['Geologists', 'get', 'excited', 'word', 'rock', 'word', 'quartz']

Balanced Text Data

Balanced number of comments from each subreddit of 0.52% /r/astronomy 0.48% /r/geology

Similar word count distribution across all the comments in each subreddit

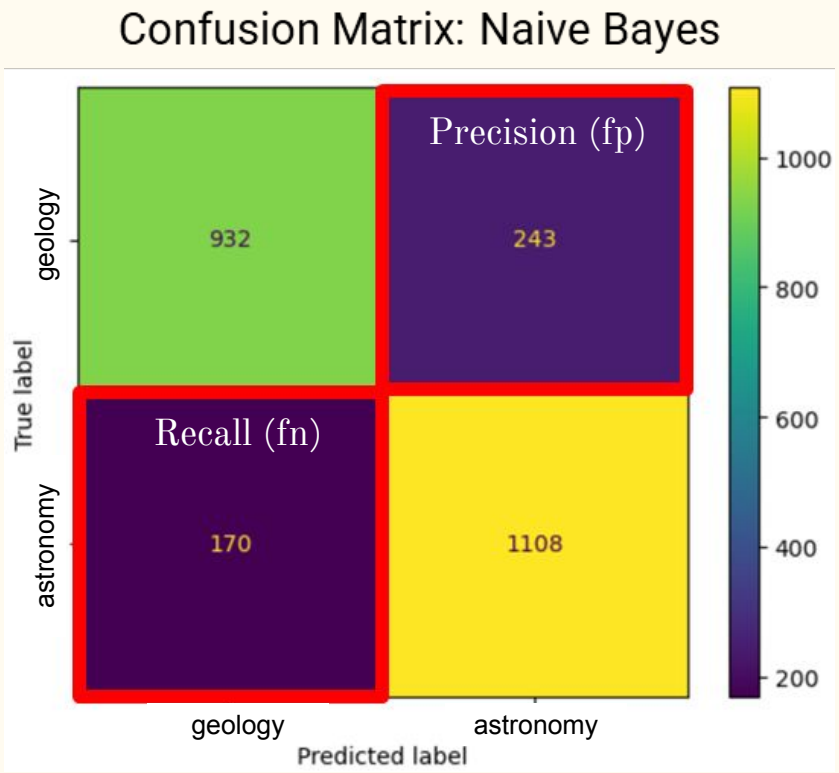


Model Selection

	Data Size	Feature Relationships	Model Complexity	Class Imbalance
Multinomial Naïve Bayes	✓		✓	✓
Random Forest	✓	✓	✓	
Our Data	Entries (10k) Features (vectorized word count: 12k)	Keywords unique to each classes	Binary Classification	52% vs 48%

Model Evaluation

	Multinomial Naive Bayes	Random Forest
Accuracy Score	**0.83**	0.80
Recall	**0.86**	0.84
Precision	**0.82**	0.79
F1 Score (consider both)	**0.84**	0.80



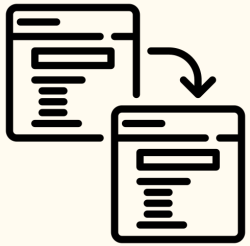
Interpretation

- **Recall**: subreddit is astronomy, but we do not know.
- **Precision**: We think it is astronomy subreddit, but it is not.

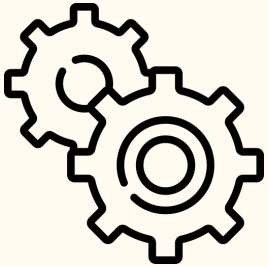
We want to avoid both cases, **F1 score** metrics can be use.

Conclusion & Recommendation

- **Accuracy 50% → 80%**
- **Production implementation**
- **Increase ads engagement**



- **Implement on other /r**
- **Implement on other platforms**



- **Feature Engineering (n_gram, weightage)**
- **Model Complexity (n_estimators, tree depth)**

Thank you!