NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING FOR THE ABORTION RIGHTS CONVERSATION

Annie Bishai

PROBLEM STATEMENT

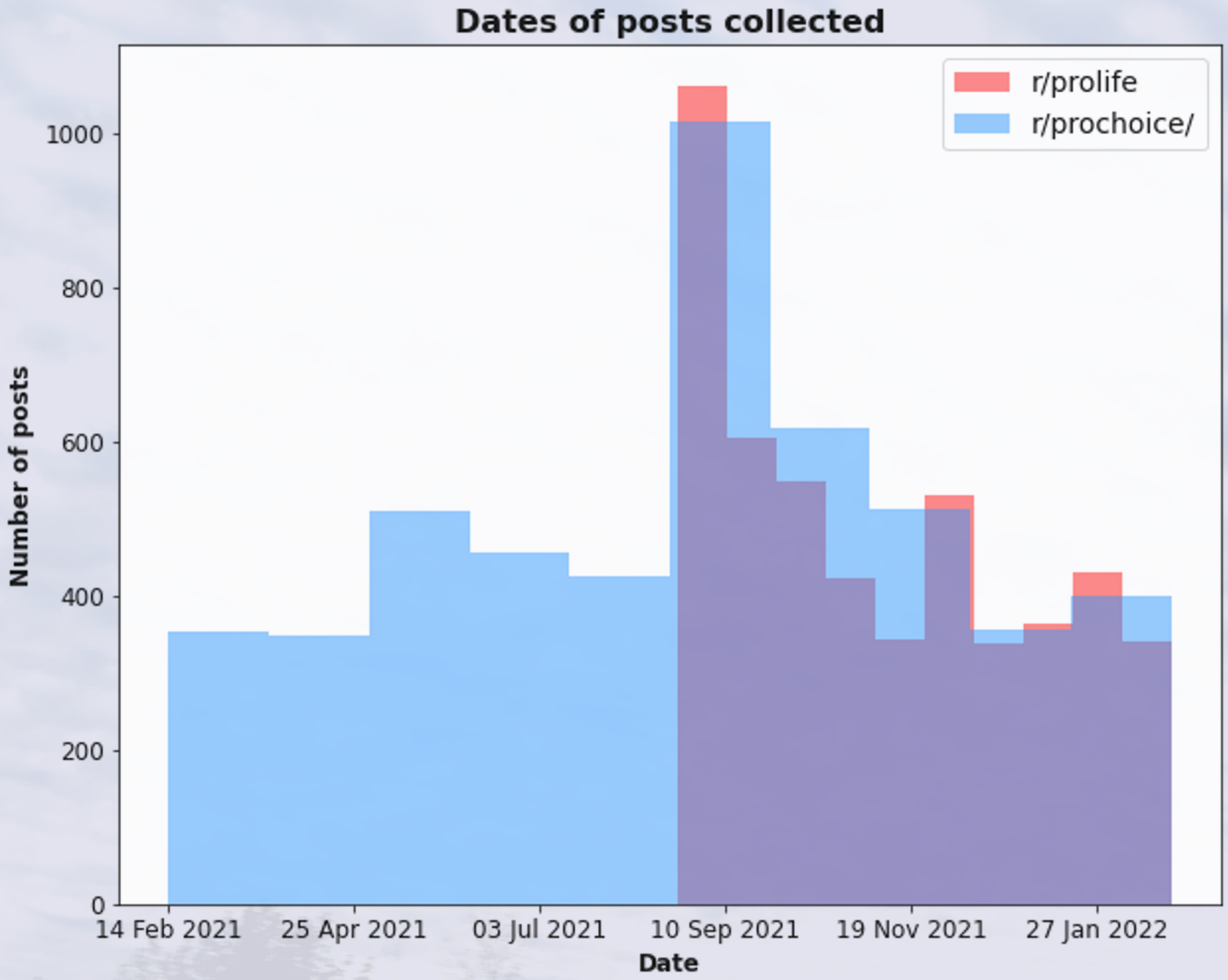
→ What is the best machine learning model to predict whether a Reddit post comes from a user who opposes abortion access or one who supports it?

→ What data features work best as predictors?

→ Are there observable language differences between posts by users in the two categories?

PROJECT OVERVIEW

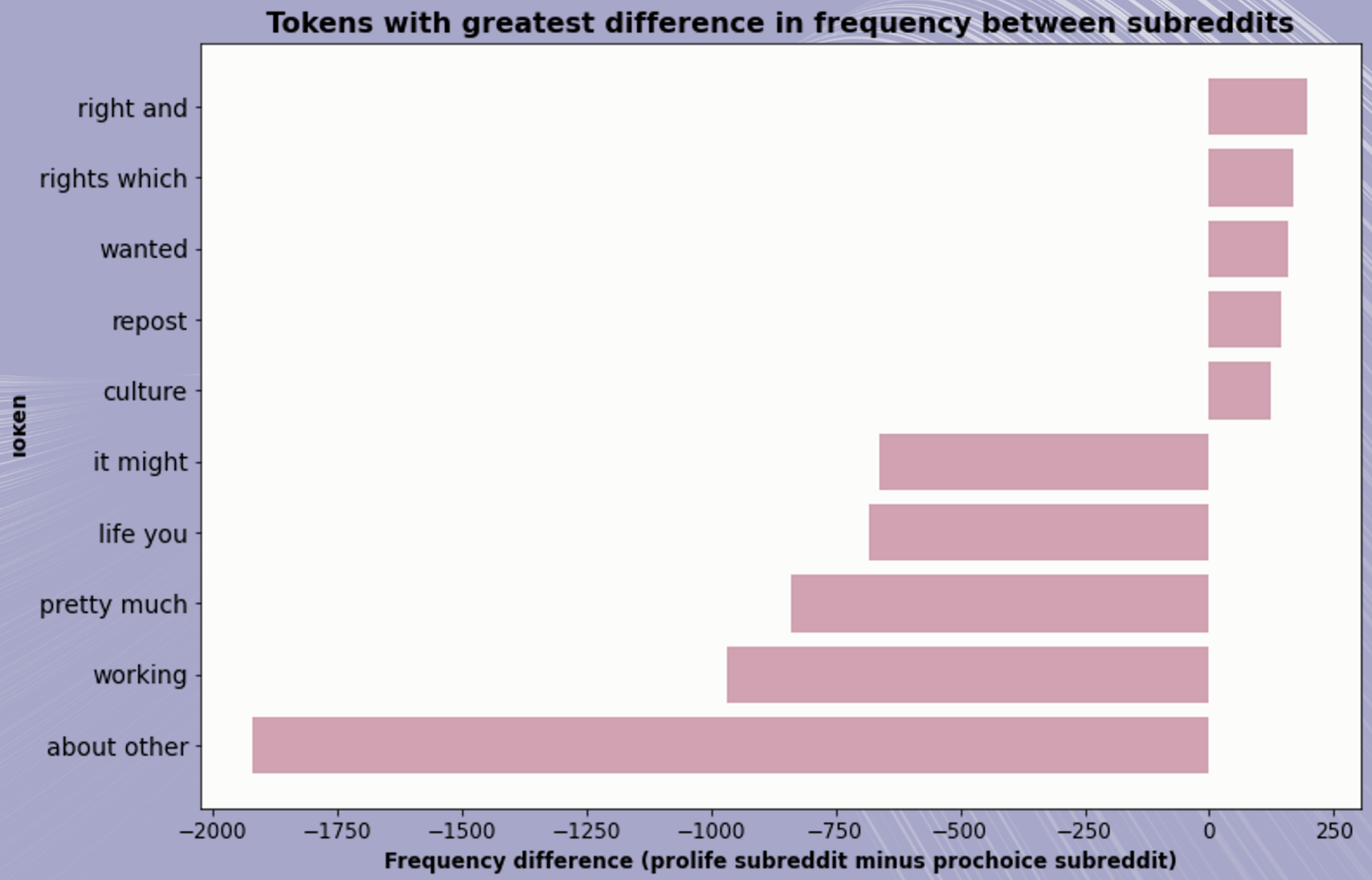
1. Data collection
2. Exploratory data analysis
3. Feature engineering
4. Modeling with several classification models
5. Evaluation of best model
6. Conclusions and steps for further research



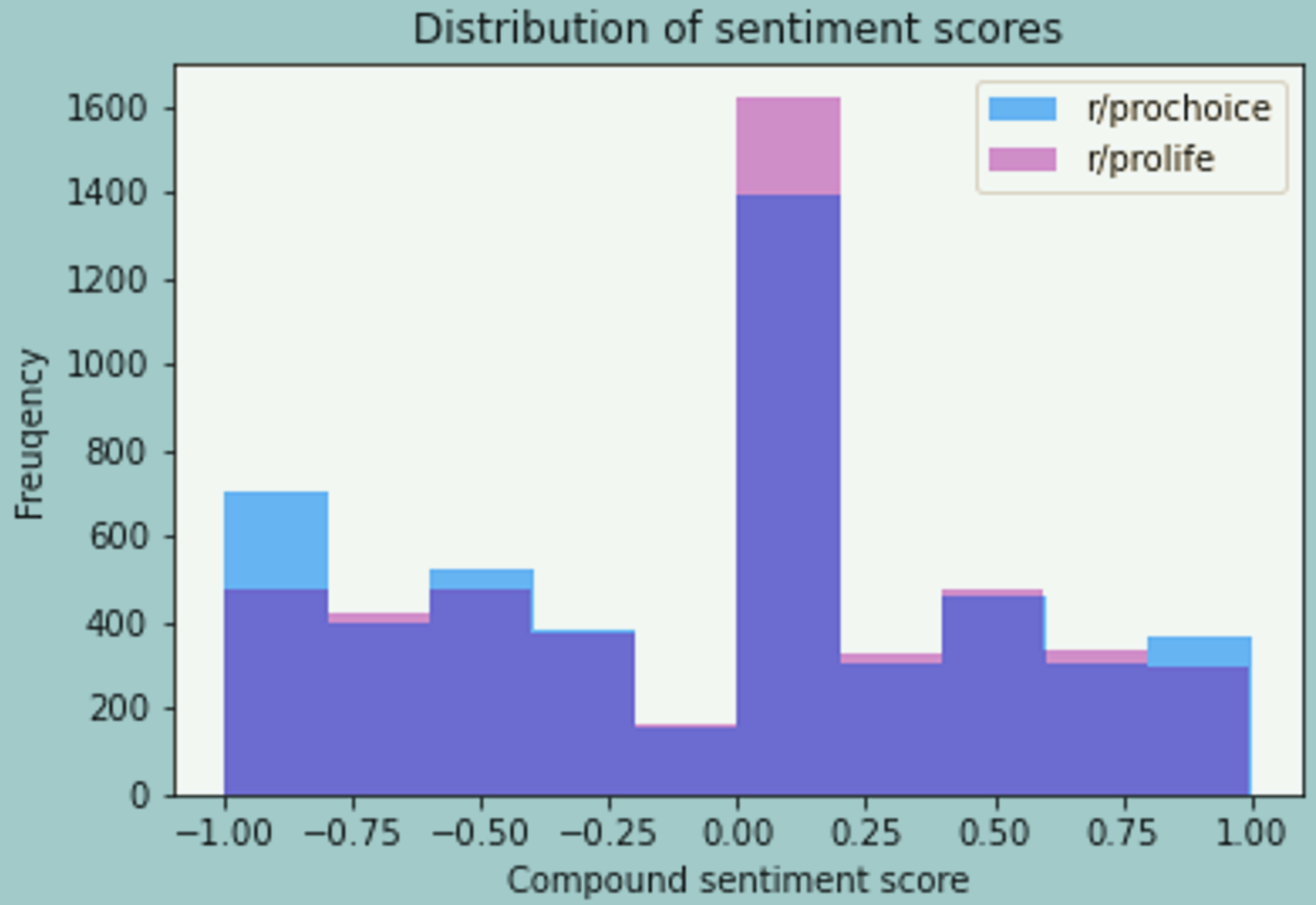
DATA

- 4,986 posts from r/prolife
- 5,000 posts from r/prochoice

WORD FREQUENCY DIFFERENCES



SENTIMENT ANALYSIS



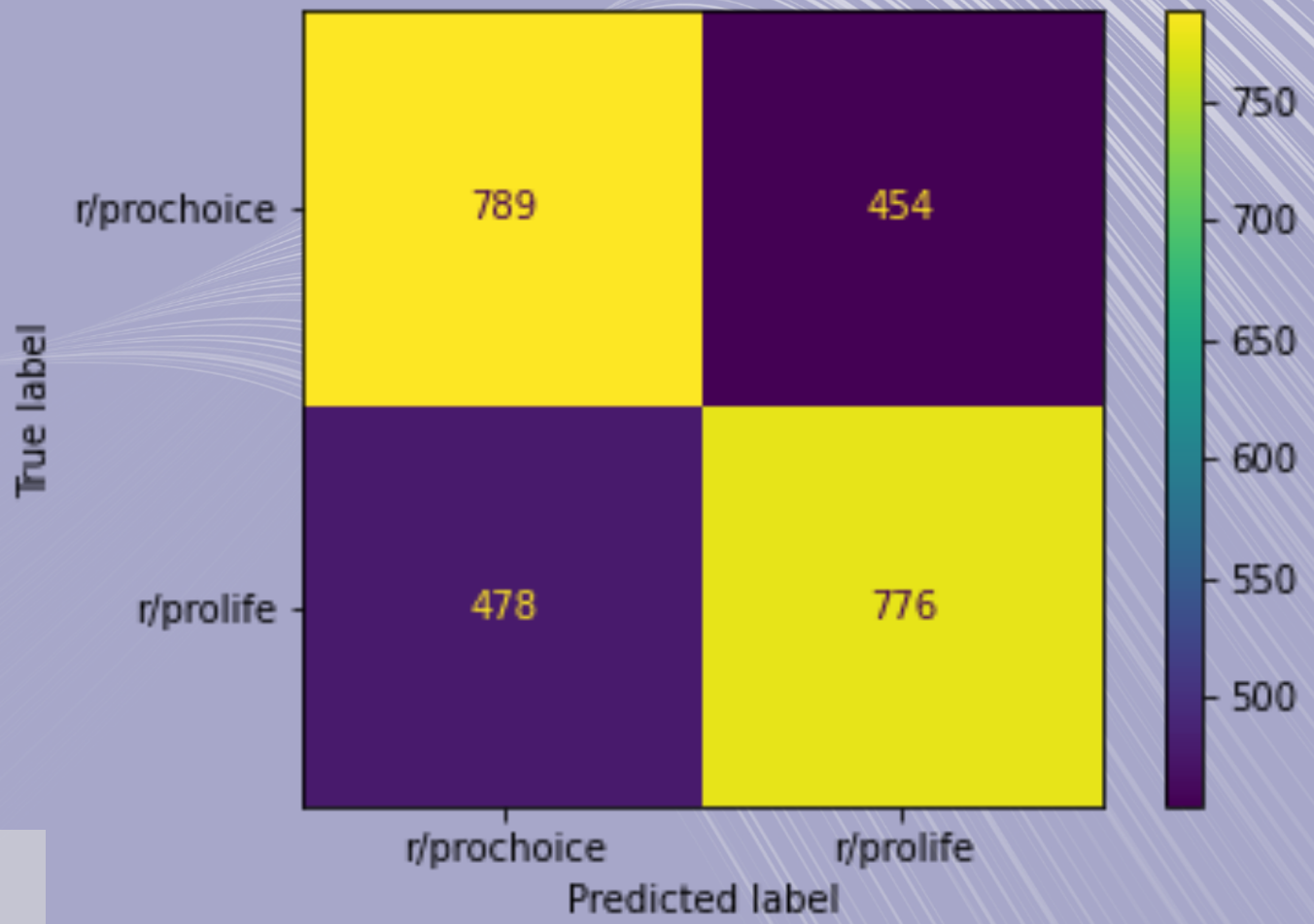
MODEL PERFORMANCE

MODEL	ACCURACY SCORES (train/test) **Vectorization only**	ACCURACY SCORES (train/test) **Vectorization and sentiment scores**
Logistic Regression	.729 / .629	
Decision Tree Classifier	.622 / .626	~.6 / ~.6
Decision Tree Classifier (TF-IDF vectorization)		.601 / .596

BASELINE ACCURACY: .500

MODEL PERFORMANCE

Decision tree
features: CountVectorizer tokens (n=12000)
n-gram range: (1, 3)
stop words: English



61.8% accurate among actual pro-life
63.0% accurate among predicted pro-life
63.4% accurate among actual pro-choice

CONCLUSIONS

- Inconclusive which model and features work best
- Decision tree classifier is promising
- Some differences and language and sentiment are observed; more research is warranted

FURTHER STEPS

- Develop more efficient modeling strategies
- Refine vectorizer to recognize more language nuances
- More data