



The Language of **Reddit**

By Charles Crocicchia



Background on the Problem

I work for a linguistic research team whose goal is to visualize patterns or trends in the rapidly developing language of the online world.

Subreddits

r/Showerthoughts

A forum where people voice their realizations or epiphanies that they have discovered independently of academic/scientific research.

r/philosophy

Where people share and discuss morals, ethics, and social observations made by prominent figures in the philosophical research community.

Our Question

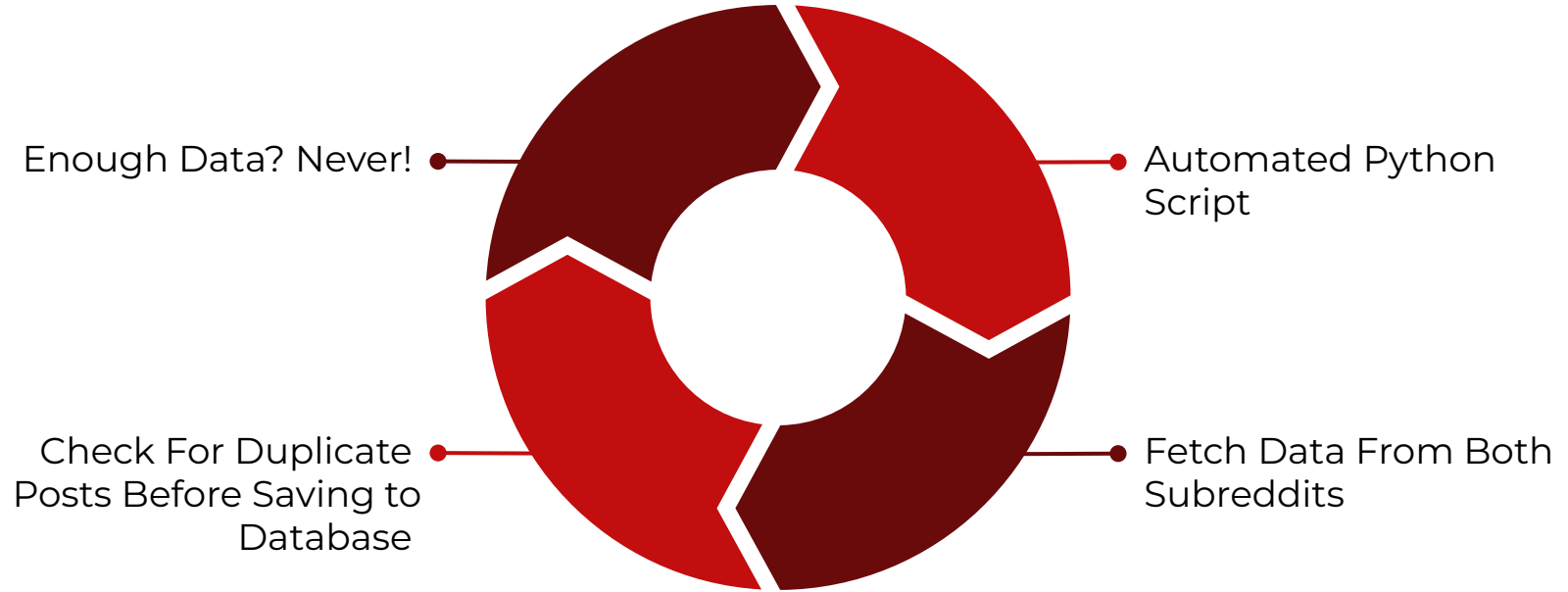
Is there a difference between our own self-made observations and those made by established philosophers?



01

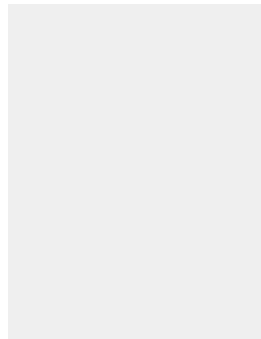
Data Collection

My Process

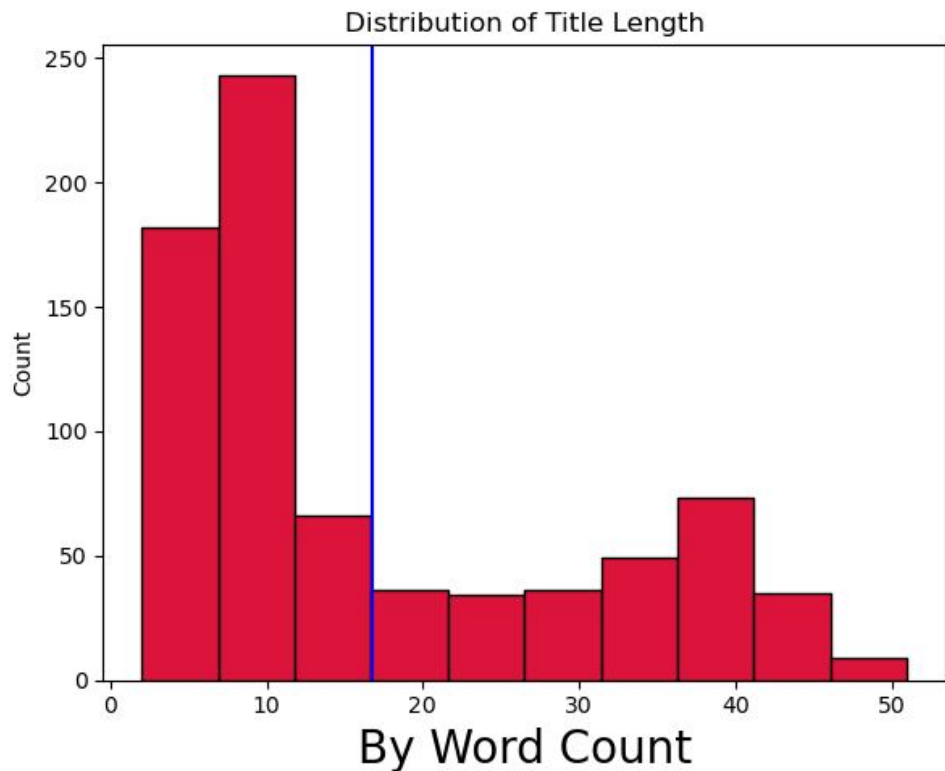
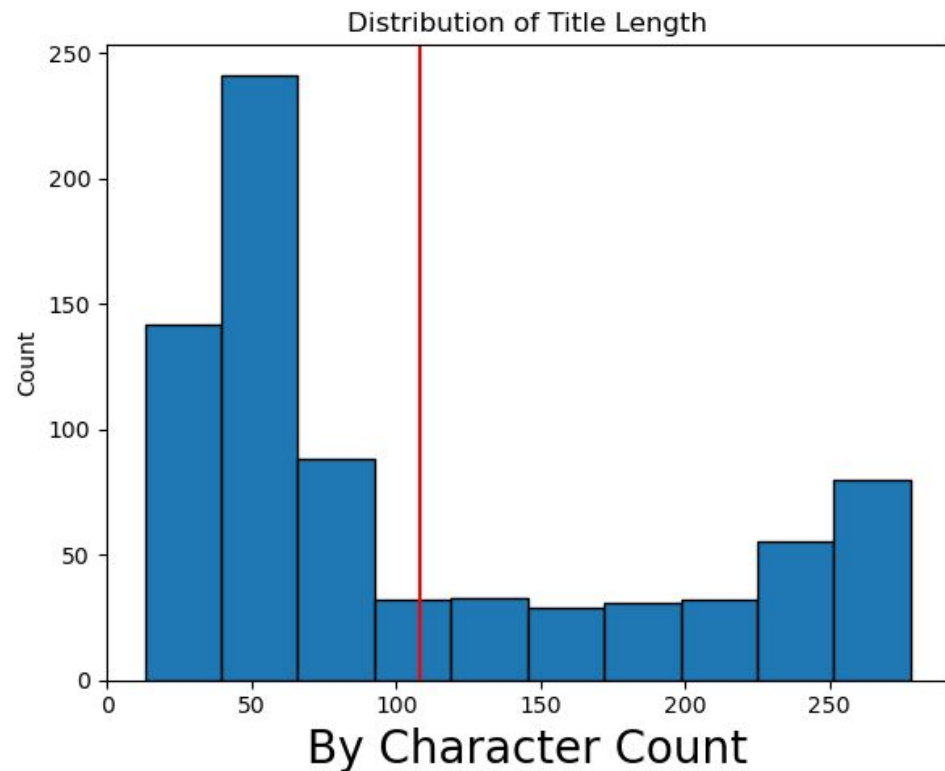


02

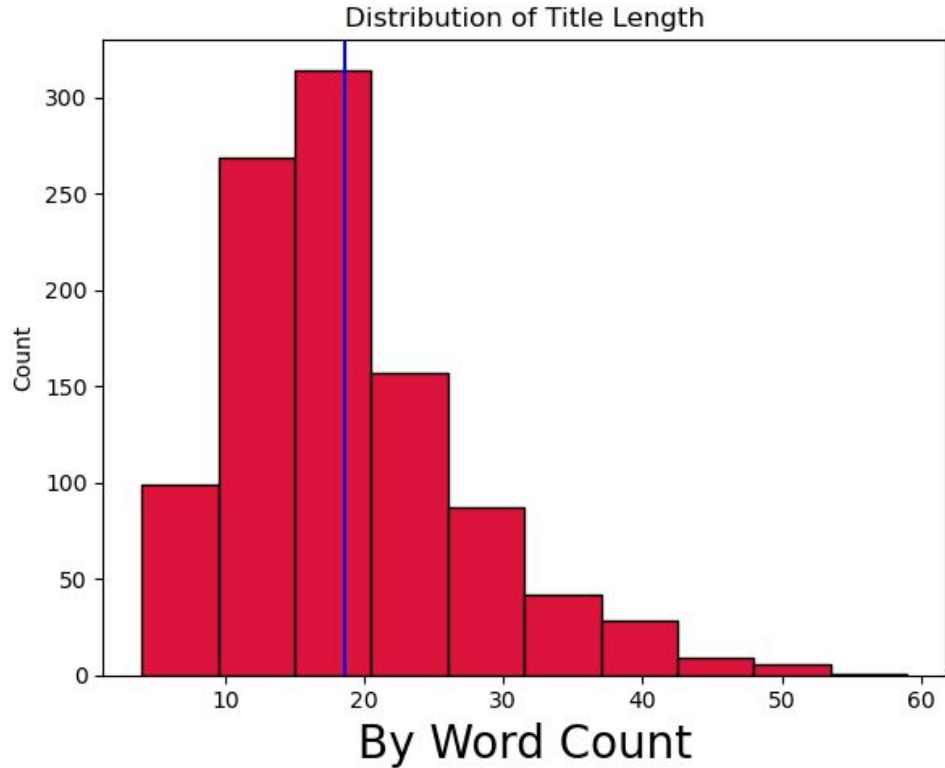
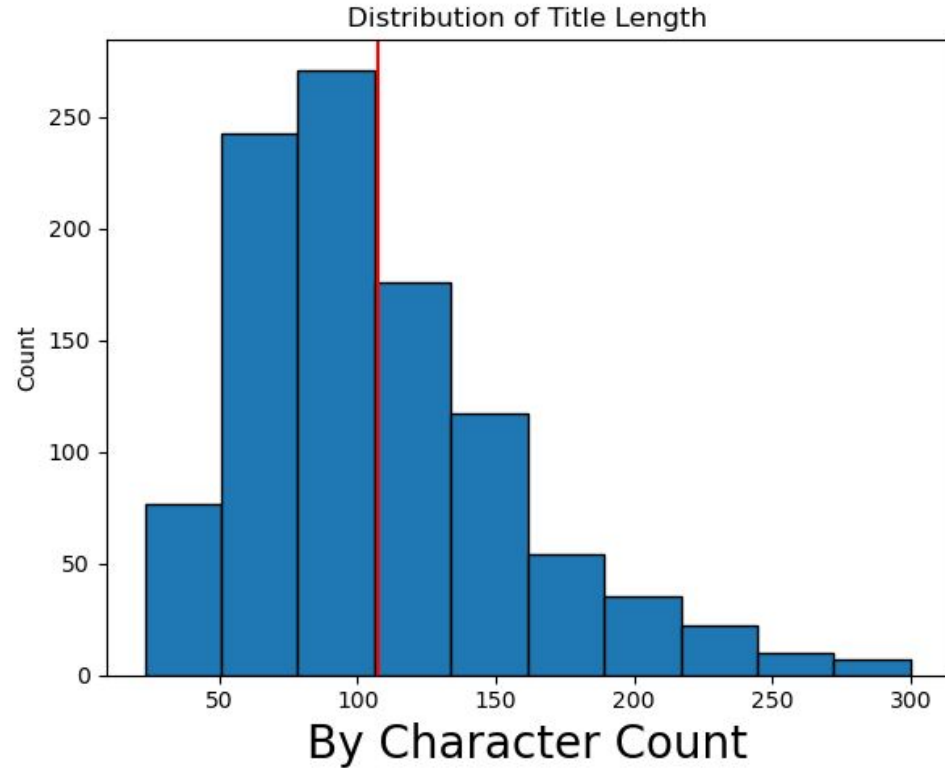
Data Analysis



r/philosophy

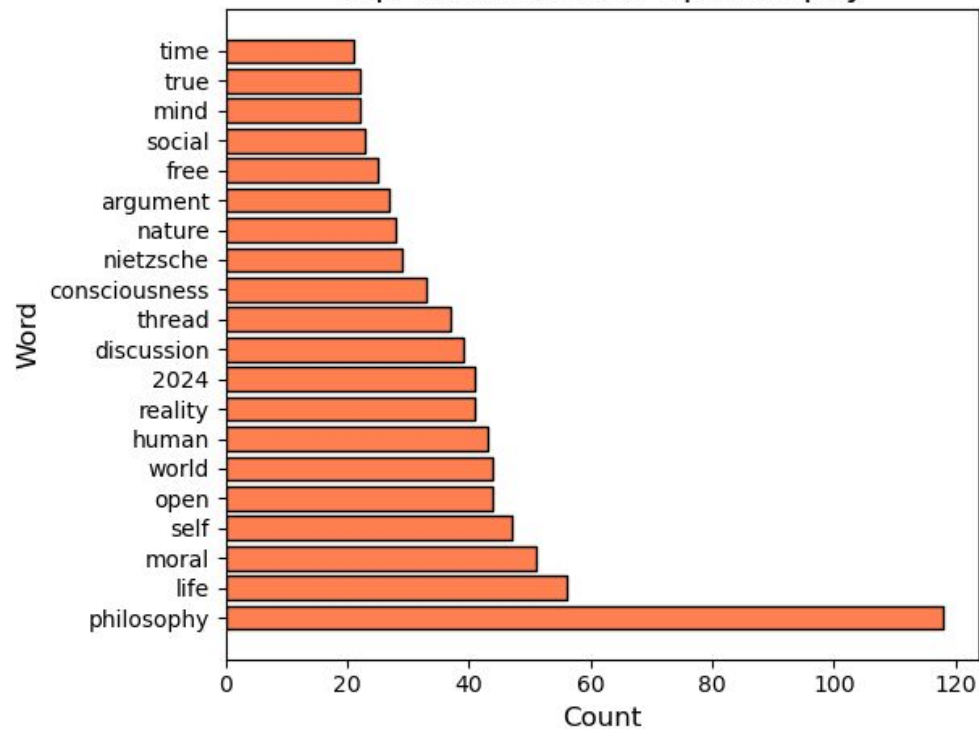


r/Showerthoughts

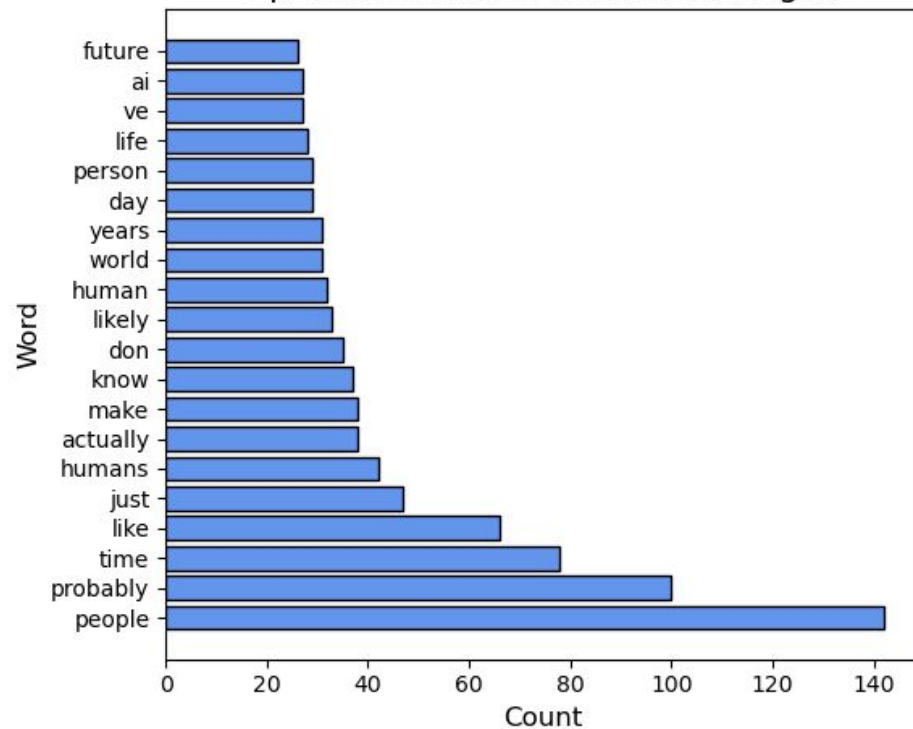


Comparing Subreddits

Top Words Used in r/philosophy



Top Words Used in r/Showerthoughts





03

Modelling

Methodology

Baseline Accuracy

57%

Models

- Logistic Regression
- Random Forest
- Naive Bayes

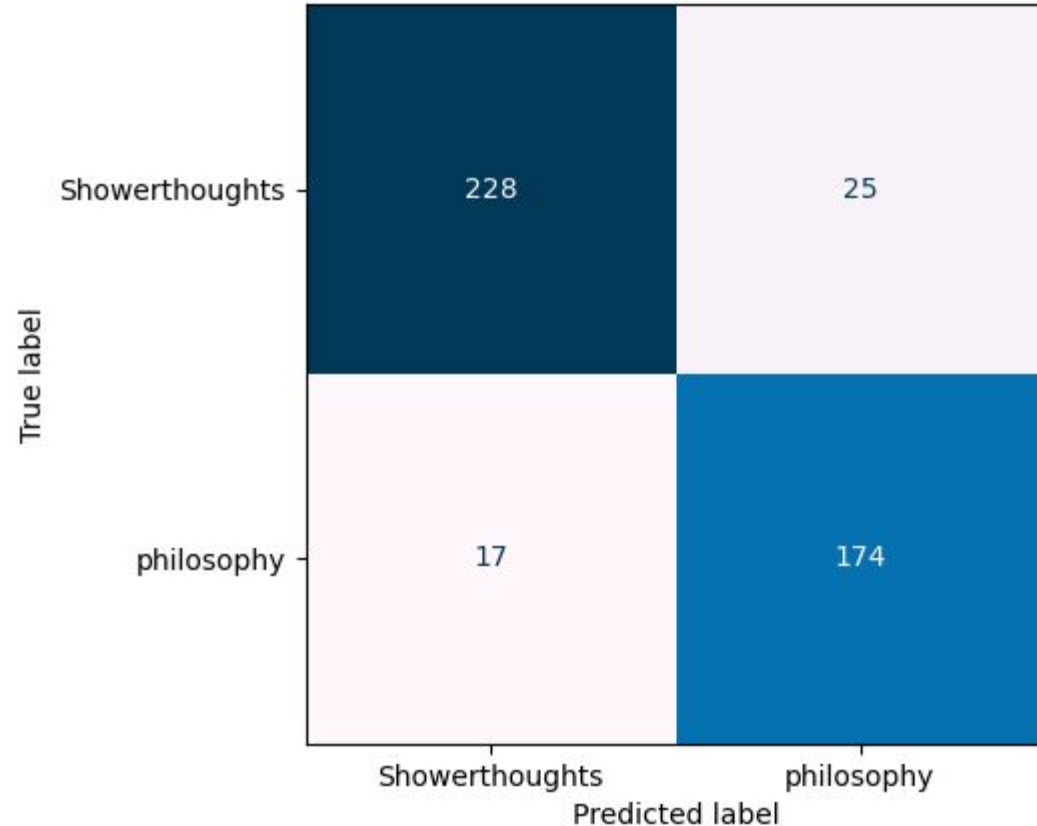
Evaluation Metrics

- Accuracy
- Bias/Variance

Trial and Error

- Tuning Parameters
- Additional EDA

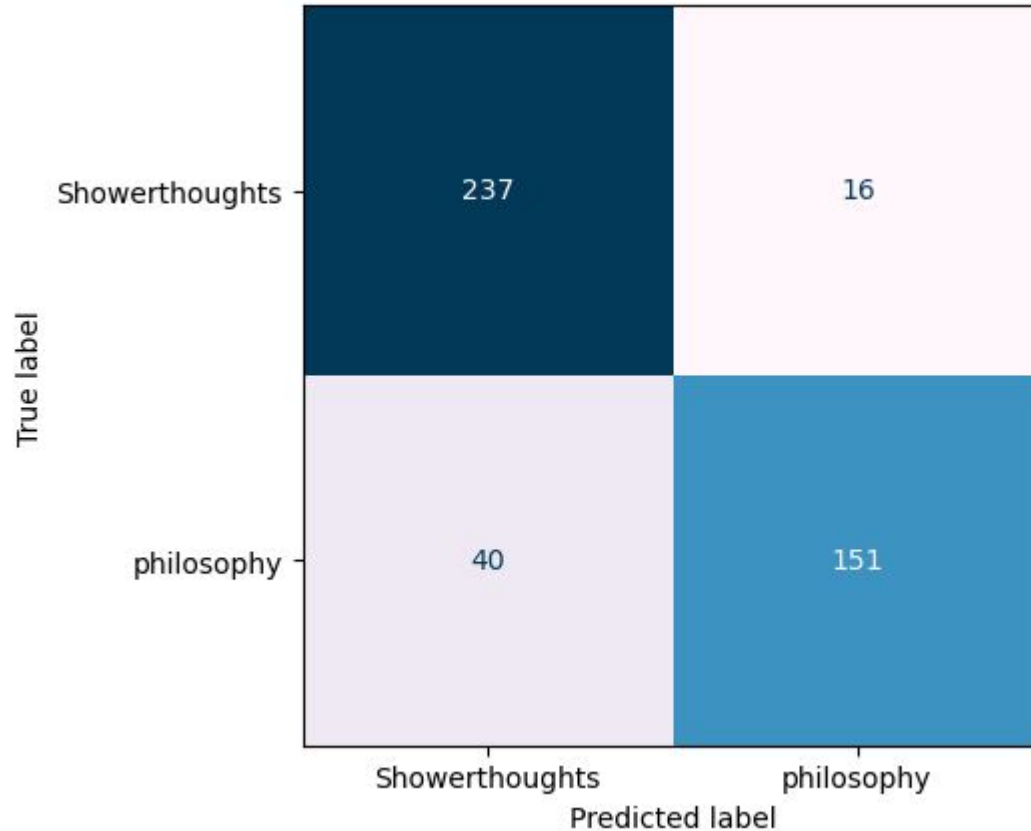
Logistic Regression



Training Accuracy
99.8%

Testing Accuracy
90.5%

Random Forest



Training Accuracy
99.8%

Testing Accuracy
87.3%

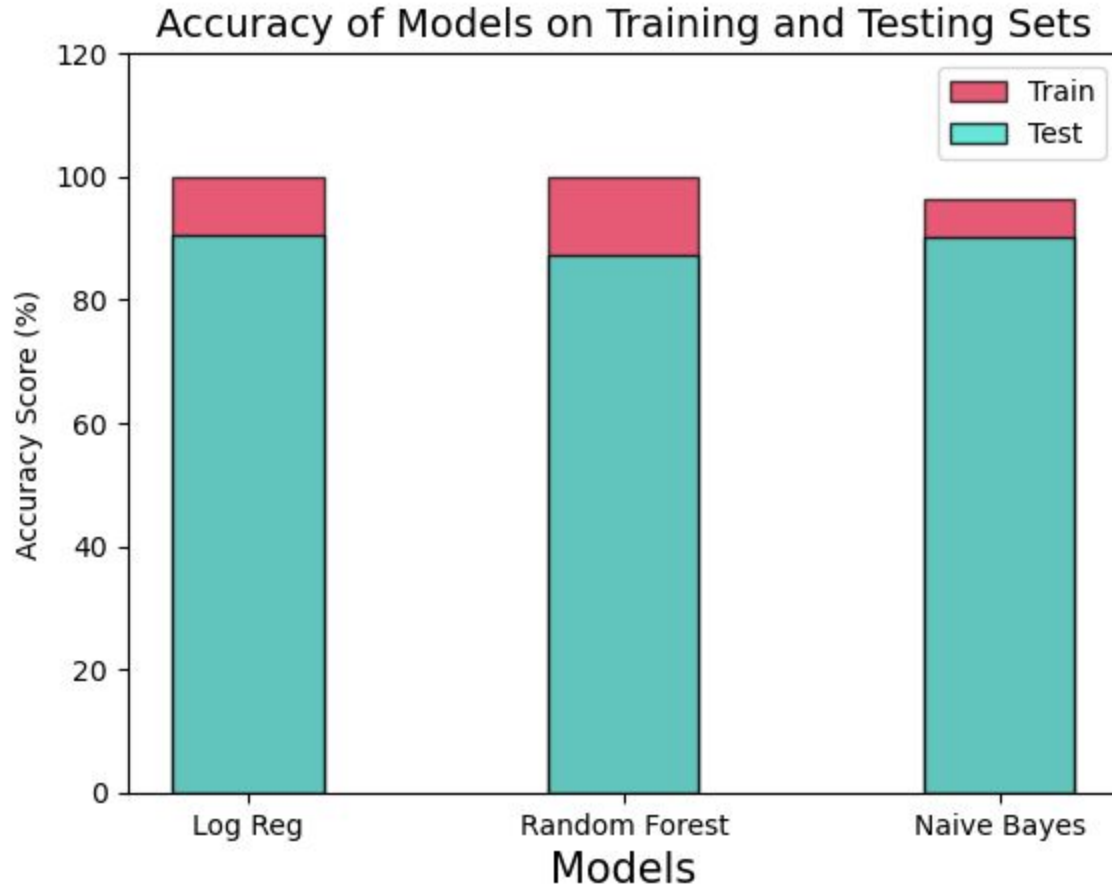
Naive Bayes

Training Accuracy
96.3%

Testing Accuracy
90.3%

True label	Predicted label	
	Showerthoughts	philosophy
Showerthoughts	244	9
philosophy	34	157

Comparison



Conclusion

01

**Data Collection
and Model
Results**

02

**Revisiting the
Problem
Statement**

03

Next Steps

04

Thank You!