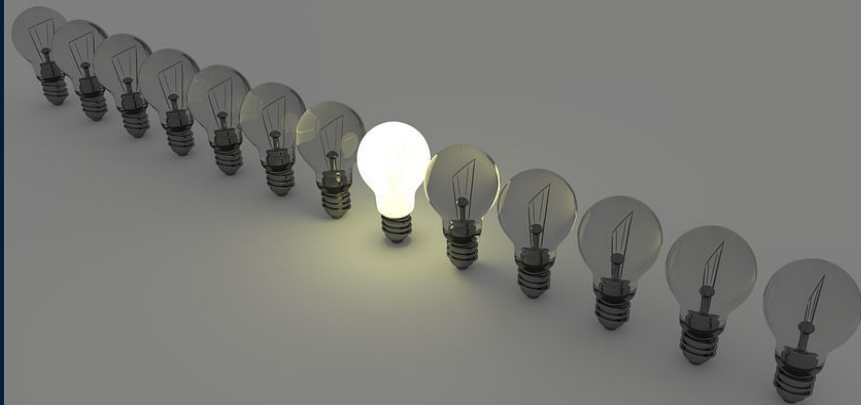# Shower Thought or Crazy Idea?

Anna Rutledge
Data Scientist

- **Problem Statement**

- **Data & Examples**

- **Analysis & Modeling**
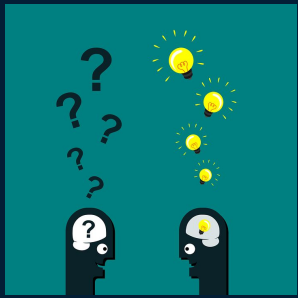
- **Results**

r/Showerthoughts

r/CrazyIdeas

"A 36-year-old has spent 18 years as an adult, so they are an adult adult."

"The entire pen industry relies on the fact that people lose their pen far more often than they use all the ink"

"All national anthems can be considered Country Music"

"Humans are the only life forms on earth that have to pay to enjoy it properly

"whenever a court undoes a law by saying its unconstitutional, the government owes reparations to everyone who that law caused problems for."

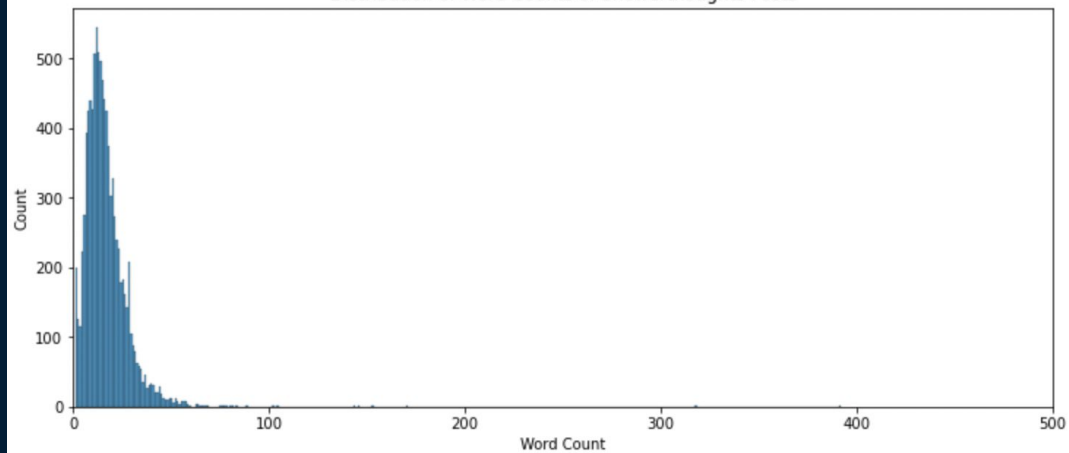"Rename the Asian and European halves of Istanbul to Eastanbul and Westanbul"

"rent out vatican city to a different religion each week"

"Clone yourself, then kill the clone, to collect your own life insurance money"
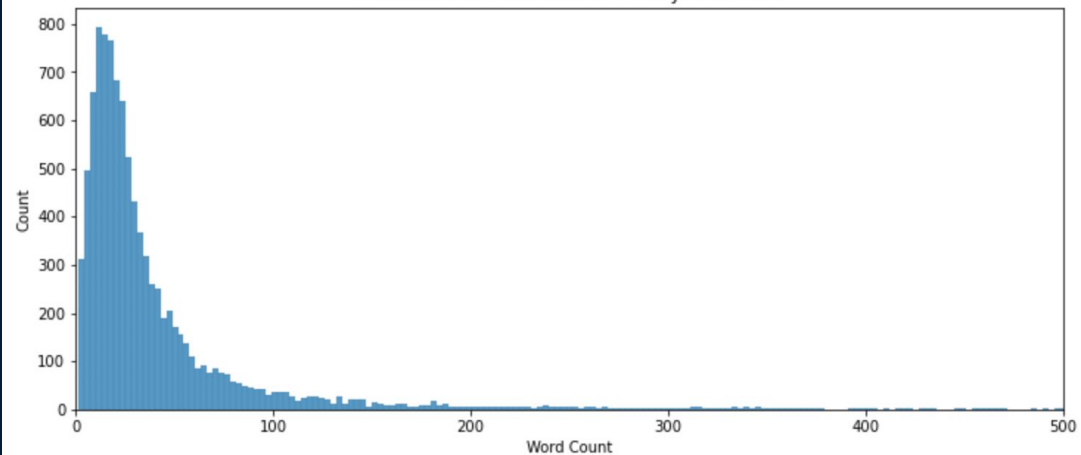
- **Text of post**

- **Word Count**

- **Letter Count**

Distribution of Word Counts of Showerthoughts Posts

Distribution of Word Counts of CrazyIdeas Posts

Null Model

- y value counts:
  - crazyideas: 9887
  - showerthoughts: 9511
- Accuracy: 0.509692
- Recall: 1.0
- Precision: 0.509692
- F1-Score: 0.675226

Models

- Naive Bayes

- KNeighbors Classifier

- Bagging Classifier

- Boosting Classifier

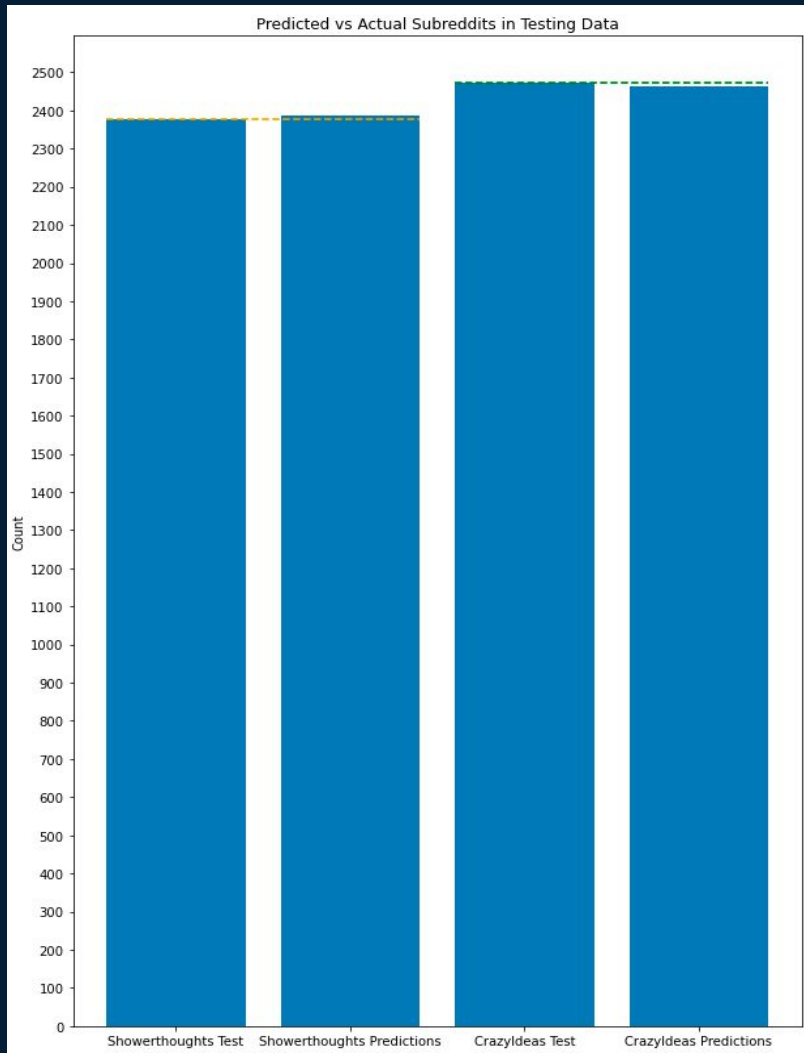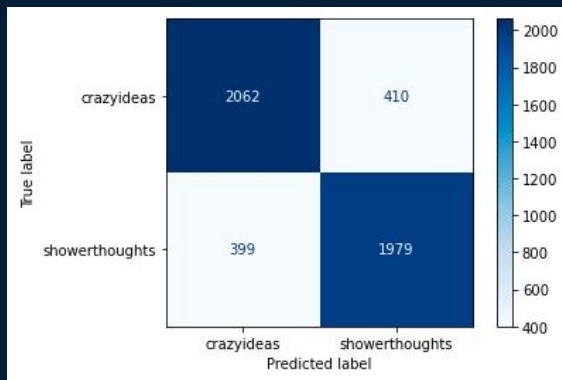- Voting Classifier: Naive Bayes & Boosting Classifier

Model	Accuracy	Recall Score	Precision Score	F1-Score
Null model	0.5097	1.0000	0.5096	0.6752
KNN Classifier – text only	0.6959	0.7512	0.6835	0.7157
KNN Classifier – all	0.7010	0.7654	0.6850	0.7230
Bagging Classifier – all	0.7408	0.7229	0.7575	0.7340
Bagging text only	0.7913	0.7872	0.8002	0.7936
Naive Bayes – all	0.7994	0.8924	0.7573	0.8193
Naive Bayes – text only	0.8054	0.8811	0.7702	0.8219
Boosting Classifier – text only	0.8099	0.7820	0.8346	0.8074
Voting Classifier – Boosting & Naive Bayes	0.8254	0.8993	0.7880	0.8400
Boosting Classifier – all	0.8322	0.8333	0.8367	0.8350

- **Optimize for accuracy**

- **Subreddit predicted with 83% accuracy**

Best Model

Boosting Classifier with 3 features:

➤ Text, Letter Count, Word Count

➤ Count Vectorizer:
  - min_df = 0, max_df = 8
  - Ngram_range = (1,2)

➤ GradientBoosting Classifier
  - learning_rate = 0.12
  - max_depth = 4
  - n_estimators = 150

Predicted vs Actual Subreddits in Testing Data

Thank you!

Questions?