

Using NLP and modeling Techniques to Determine the Source of Reddit Posts

By: Marta Fuentes-Filp

The subreddits

Am I the Asshole?

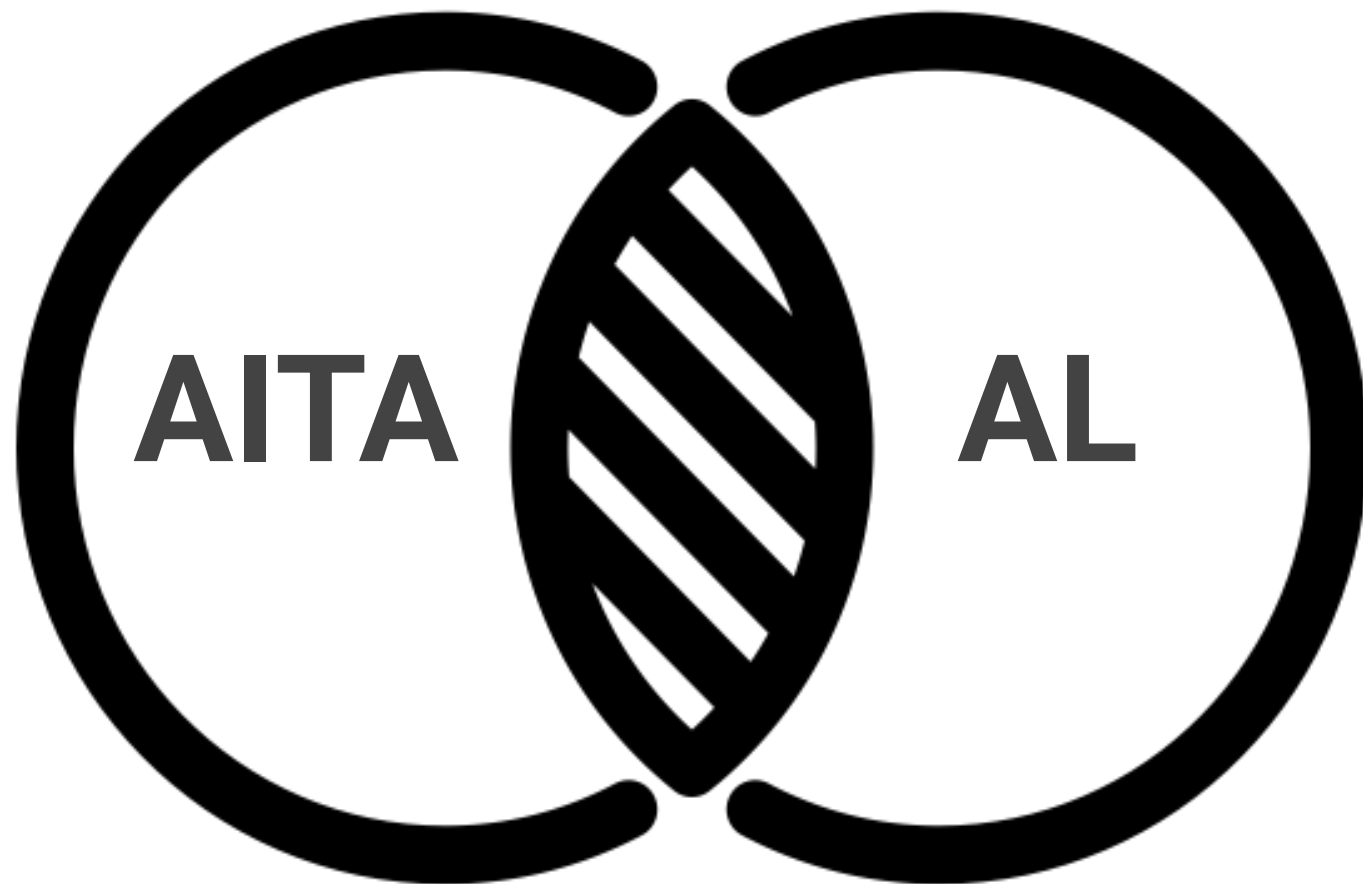
(AITA)

for bringing my sister
in law's wallet to the
restaurant when she
conveniently always
forgets it...?

Ask Lawyers

(AL)

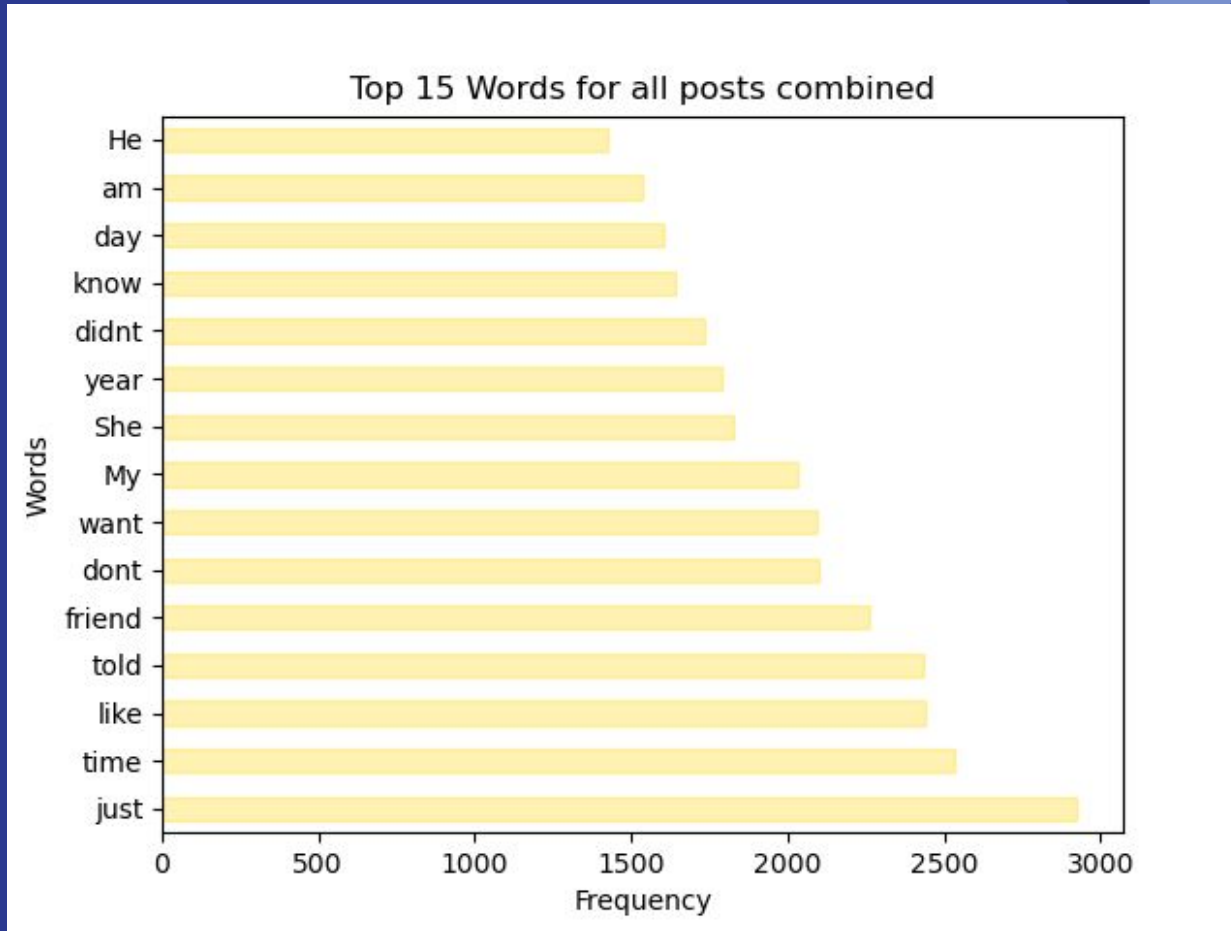
Assume somebody
found buried treasure on
someone else's land,
and started dressing up
as a ghost to scare
people ...



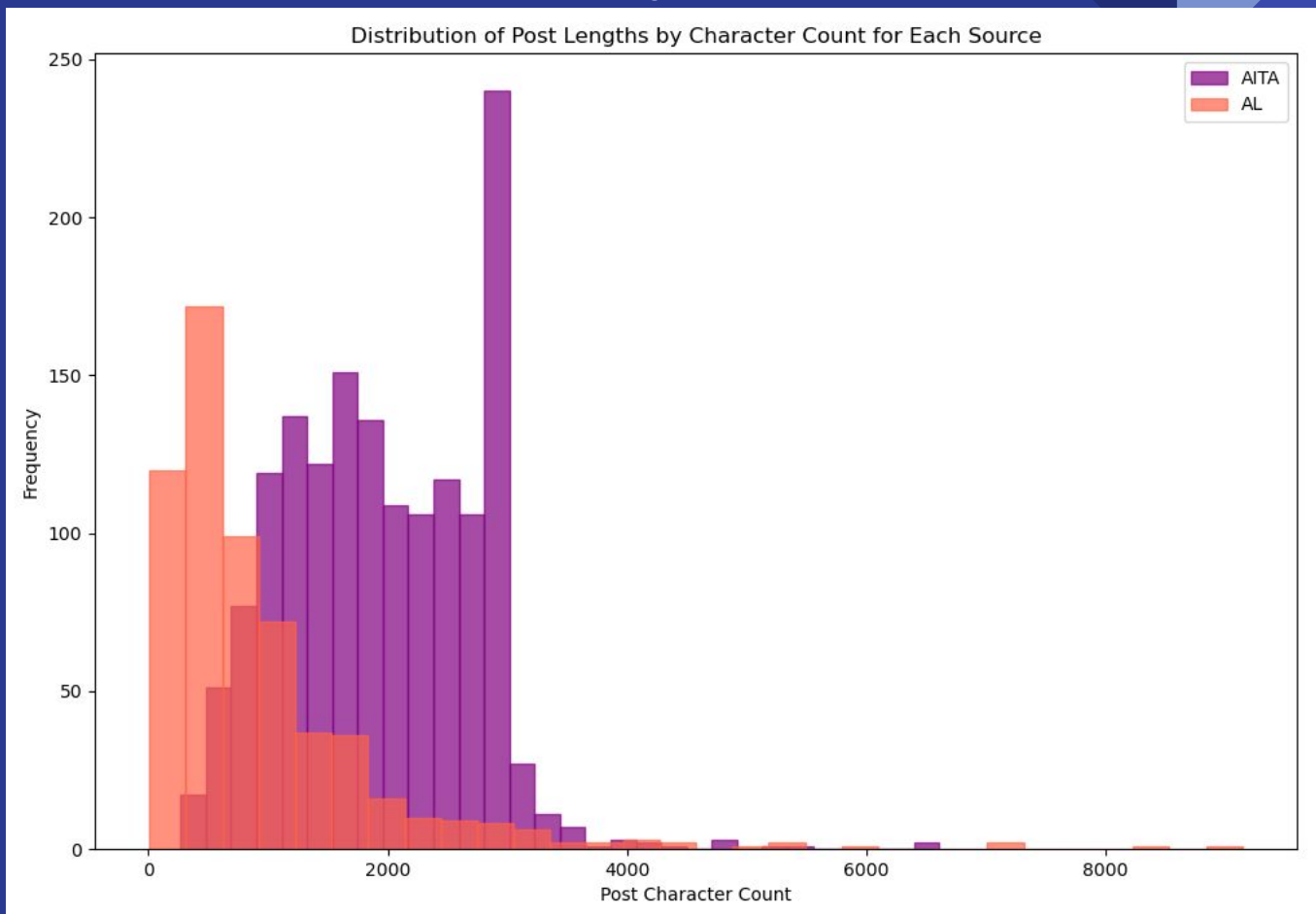
Data Preparation



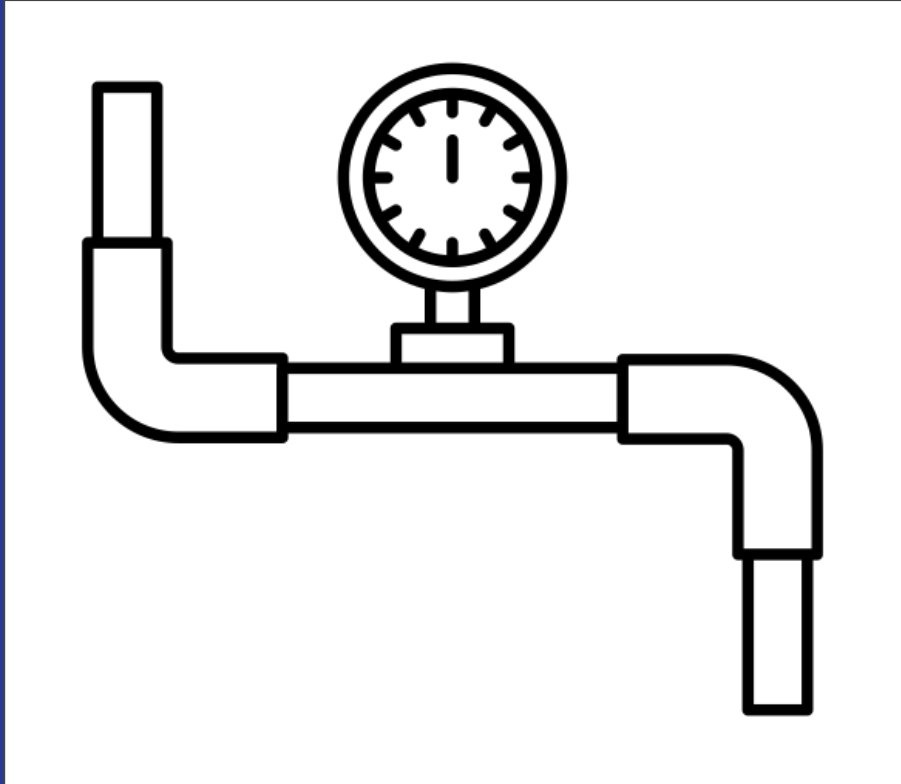
Text Processing & Analysis



Text Processing & Analysis



Text Processing & Analysis



Natural Language
Processing
+
Classification
Model
+
Hyperparameter
Optimization

Inferential insights from my Logistic Regression model:

- Negative coefficients suggest words related to...

Coefficient	Word
-2.248	court
-2.042	legal
-1.862	lawyer
-1.796	Is
-1.750	case
-1.518	Can
-1.478	property
-1.438	police
-1.351	any
-1.323	What

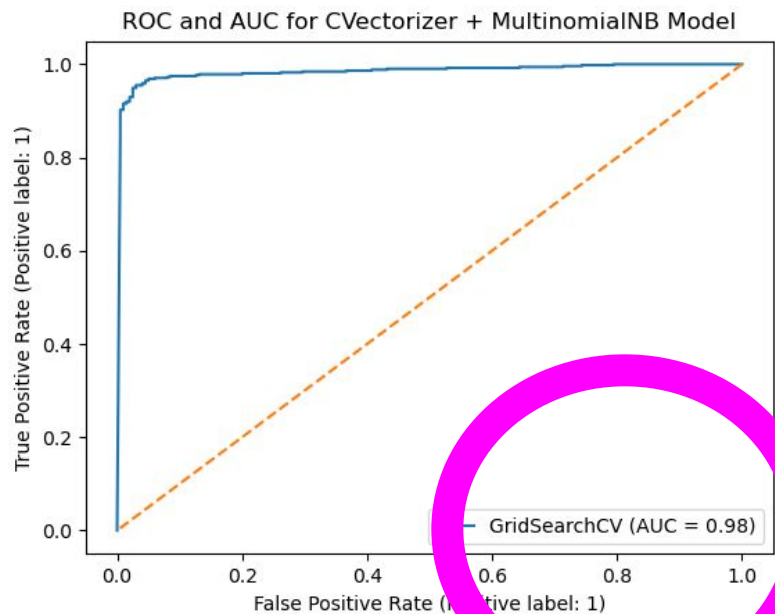
Metrics

- **Baseline Accuracy: 72%**
- **Best Accuracy Score on Testing Data: 95%**

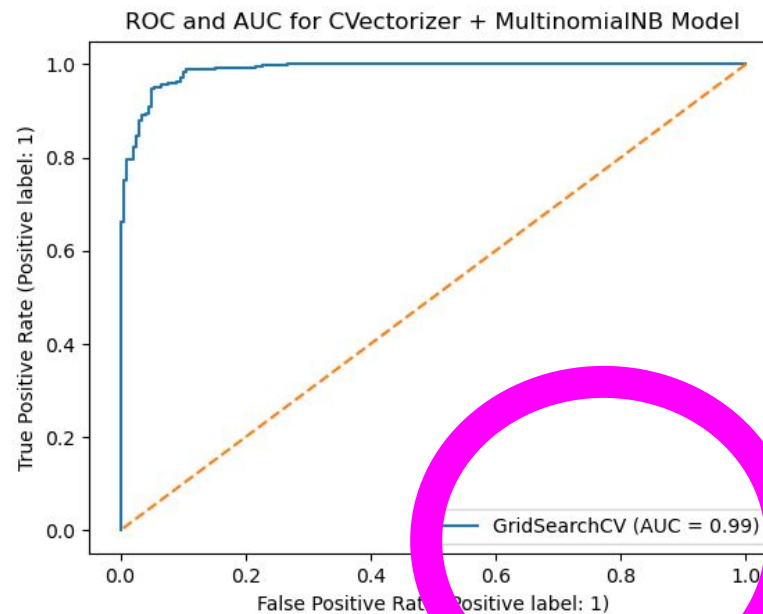
Metrics

Metric	CountVectorizer + MultinomialNB	TfidfVectorizer + LogisticRegression
Specificity	0.889	0.834
Recall	0.975	0.992
Precision	0.958	0.939
F1 Score	0.966	0.965

MODEL 1



MODEL 2



Next Steps...

More Data

Custom
Pre-processors

SpaCy & Friends

- What was the effect of the custom preprocessor used?

Beyond:

- **CountVectorizer and TFIDFVectorizer...**
- **MultinomialNB and LogisticRegression**

Thank you

What
Questions
Do
You
Have?
