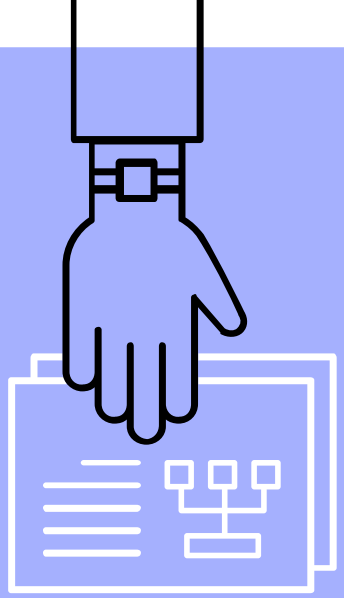
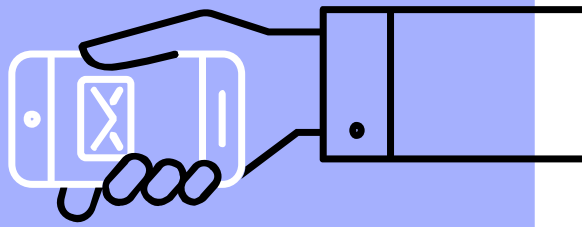
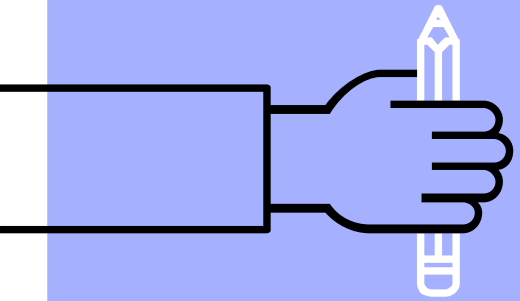
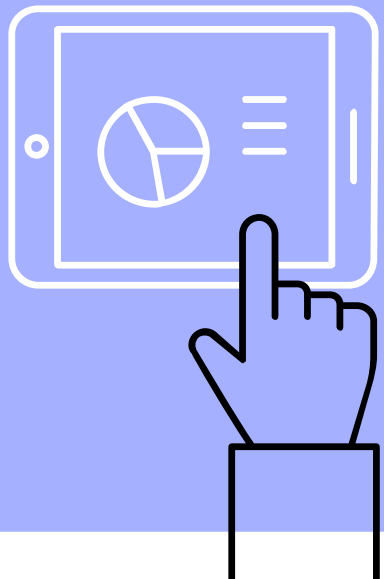


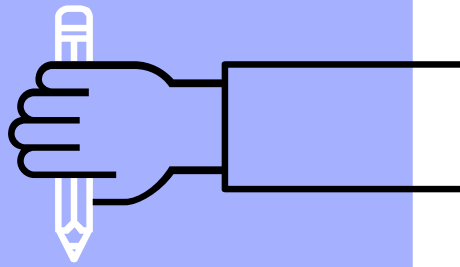


Reddit Web Scraping and NLP

Claire Hester - DSI

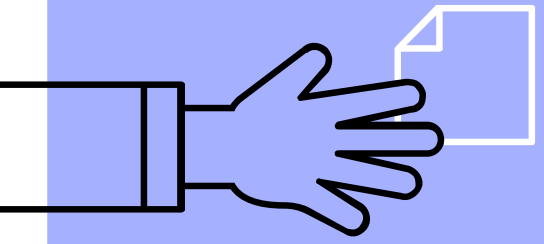


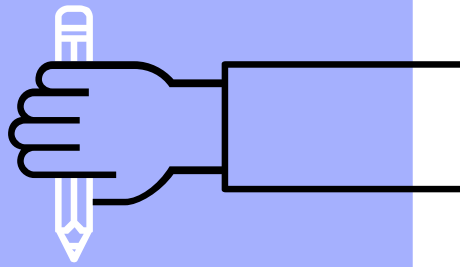




The Problem:

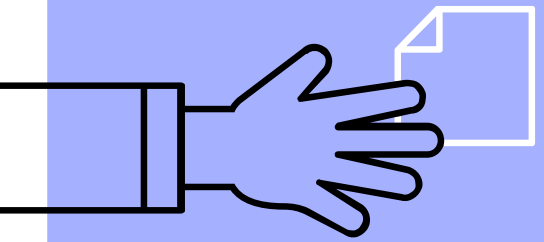
Where does our question come from - r/legaladvice or r/NoStupidQuestions?





...but First!

A quick quiz



Would it be legal if I stated in my will that I want a kahoot at my funeral and whoever wins gets my properties?

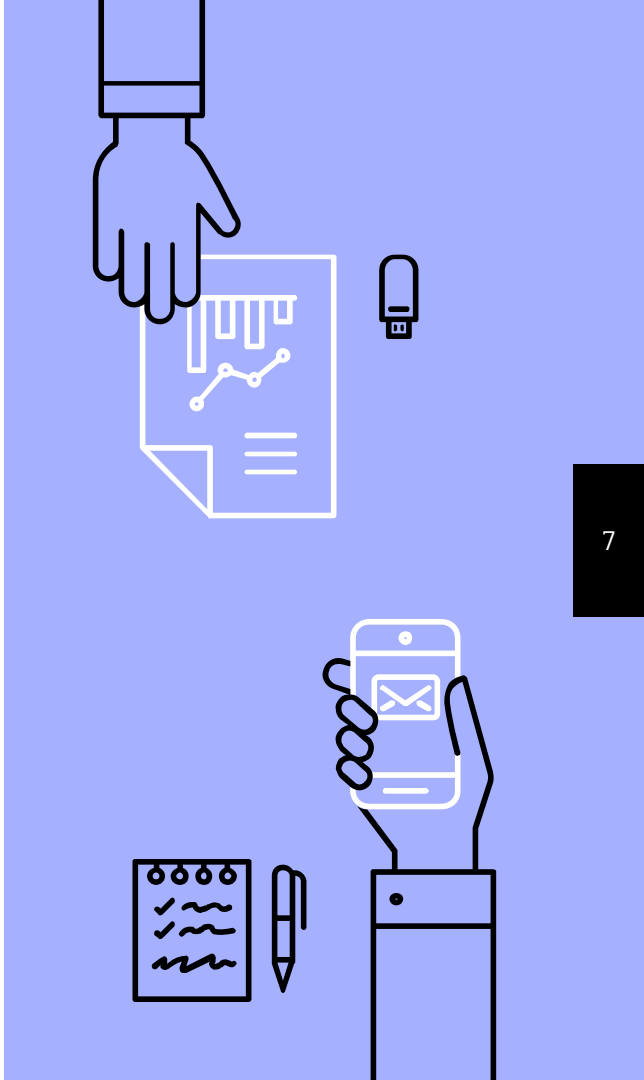
My friends have been buying fast food to throw it at people

Could a dog be a child's legal guardian under any circumstance?

Question 1: No Stupid Questions

Question 2: Legal Advice

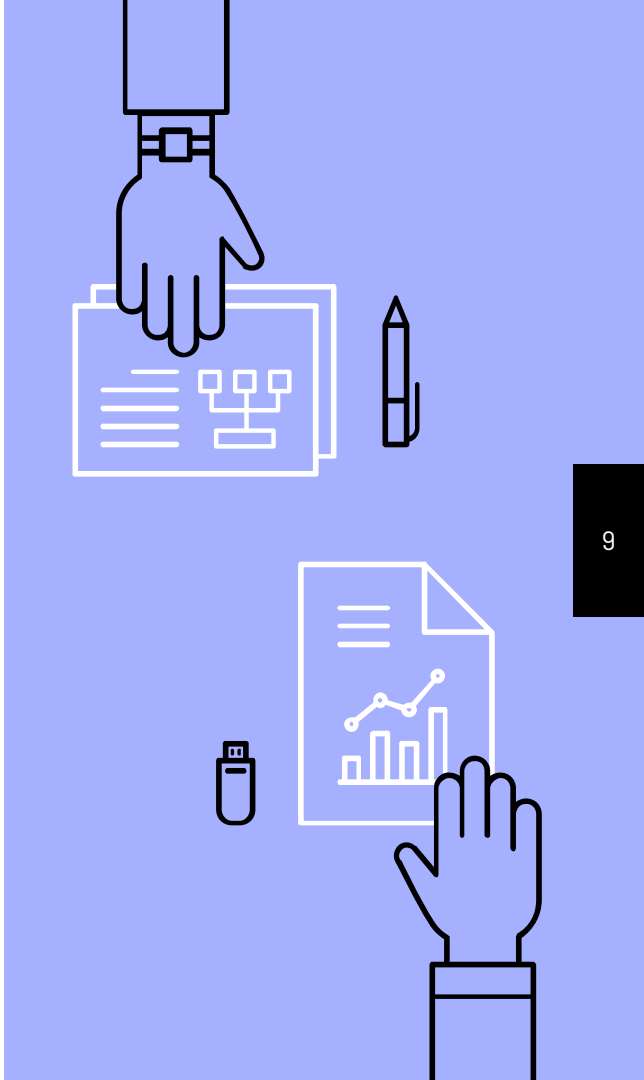
Question 3: No Stupid Questions



DATA COLLECTION and CLEANING

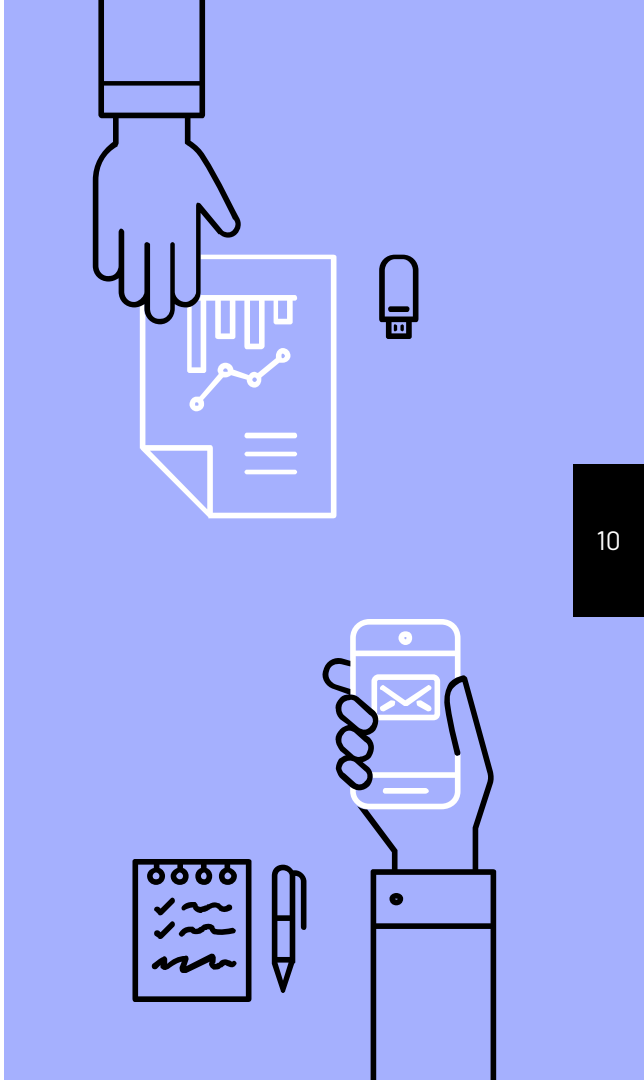
Data and Methodology

- Pulled 60,000 total submissions
- Cleaned the data
- Exploratory Data Analysis
- Compared two models
- Results



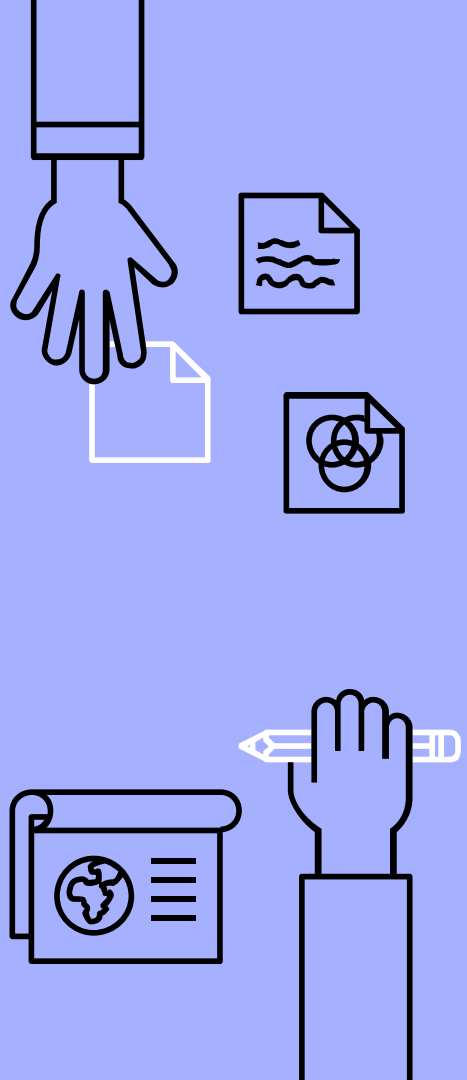
DATA COLLECTION

	Legal Advice	No Stupid Questions	Total
Submissions	28,726	26,987	55,713
Mean Word Count	220	55	140
Earliest Pull	Monday, August 17, 2020	Wednesday, September 9, 2020	n/a



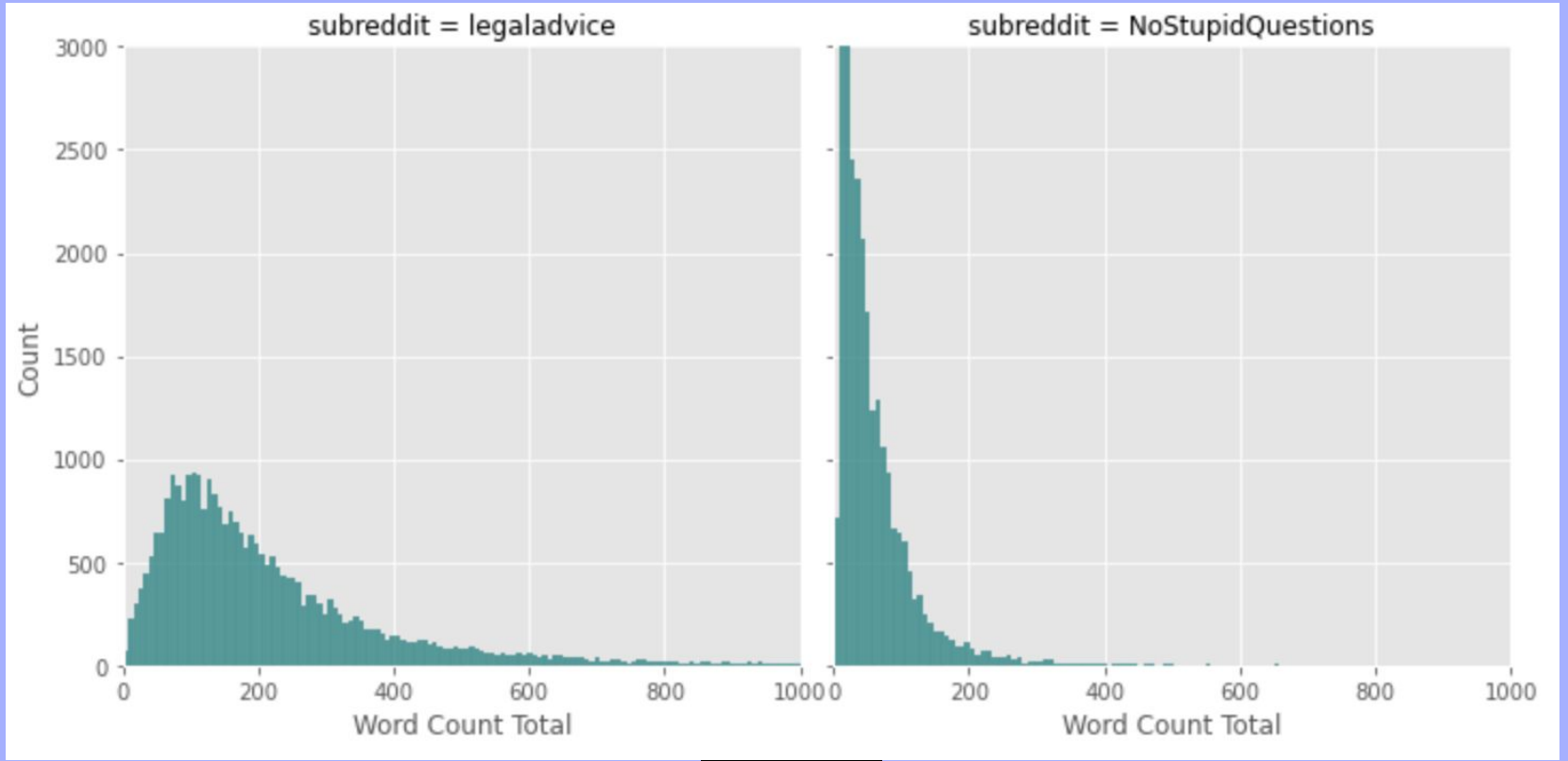
Data Cleaning

- Removed duplicate titles
- Removed submissions with [‘removed’]
- Imputed empty values with ‘blank’
- Cleaned titles and text with RegEx
- Feature engineered word count columns

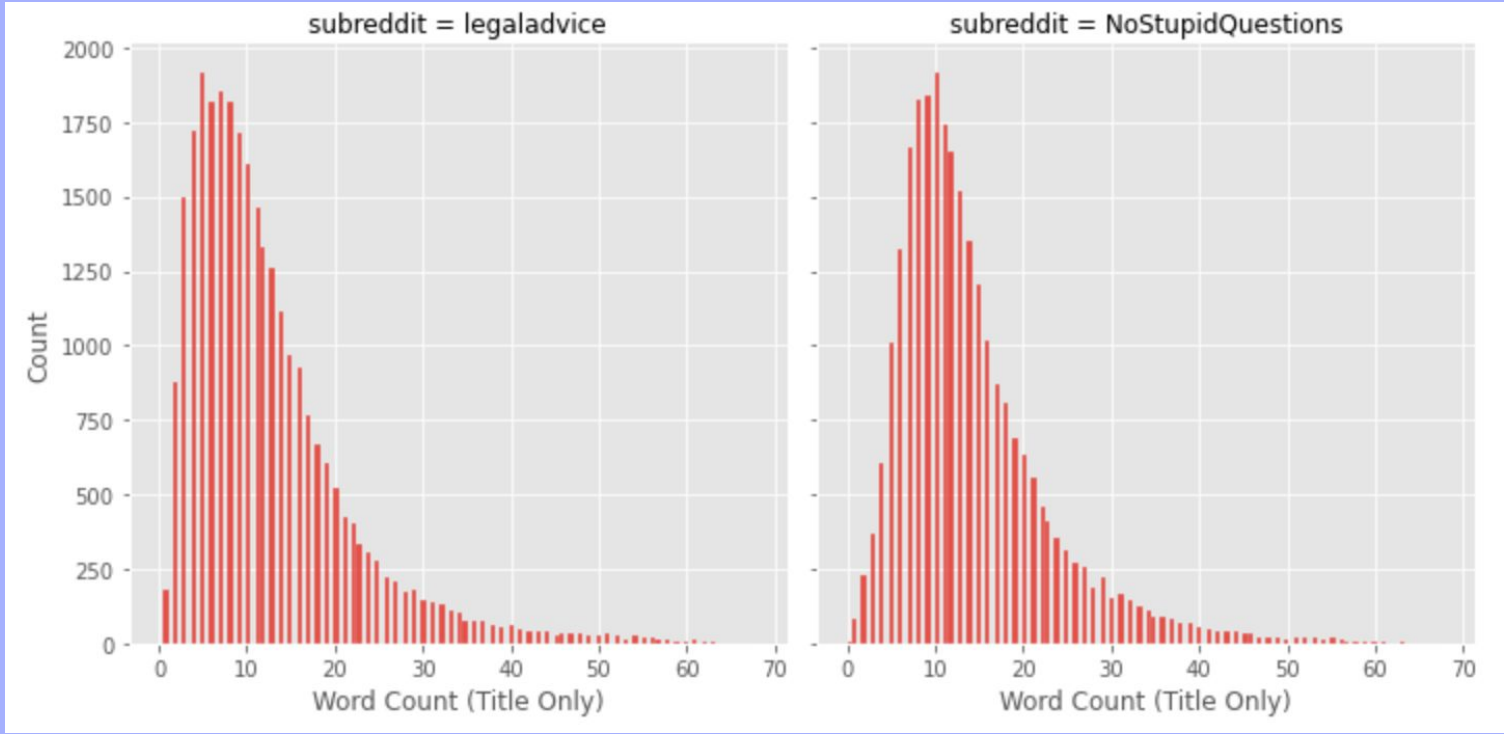


EXPLORATORY DATA ANALYSIS

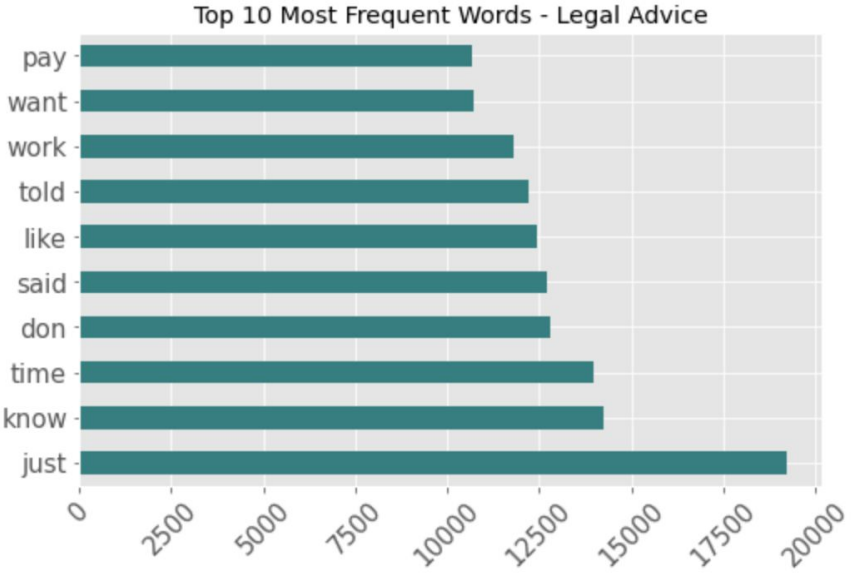
TOTAL WORD COUNT

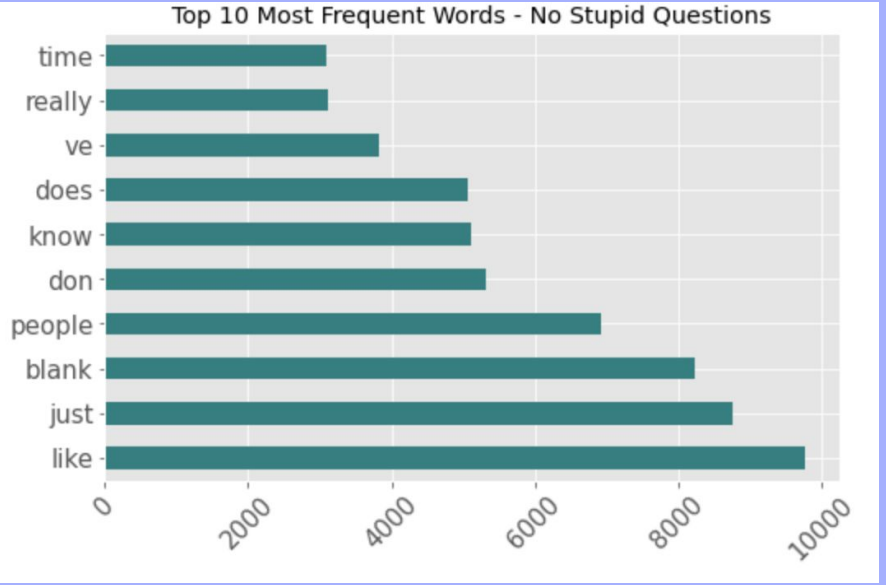


WORD COUNT - TITLE ONLY

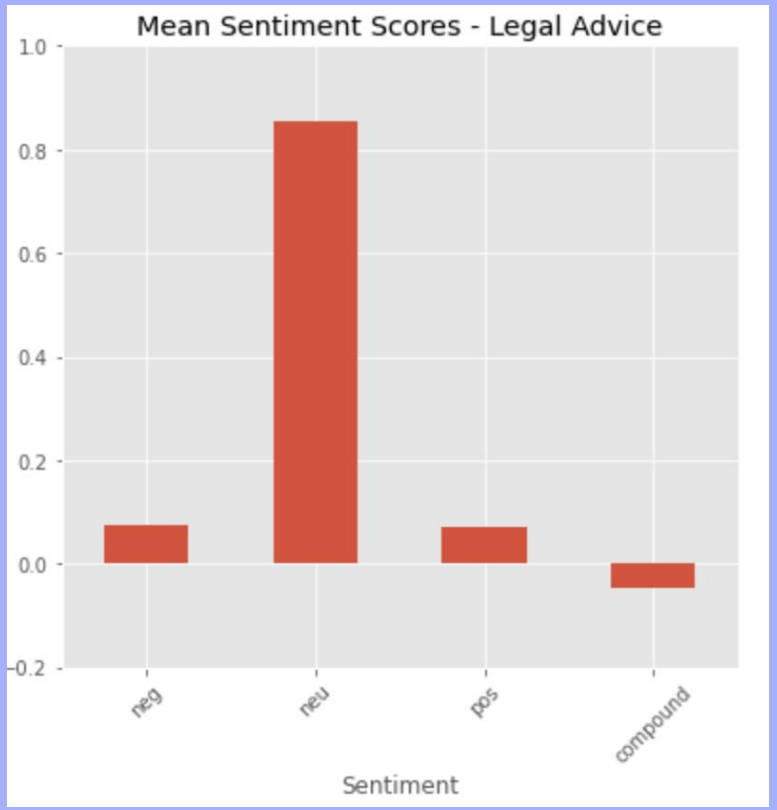


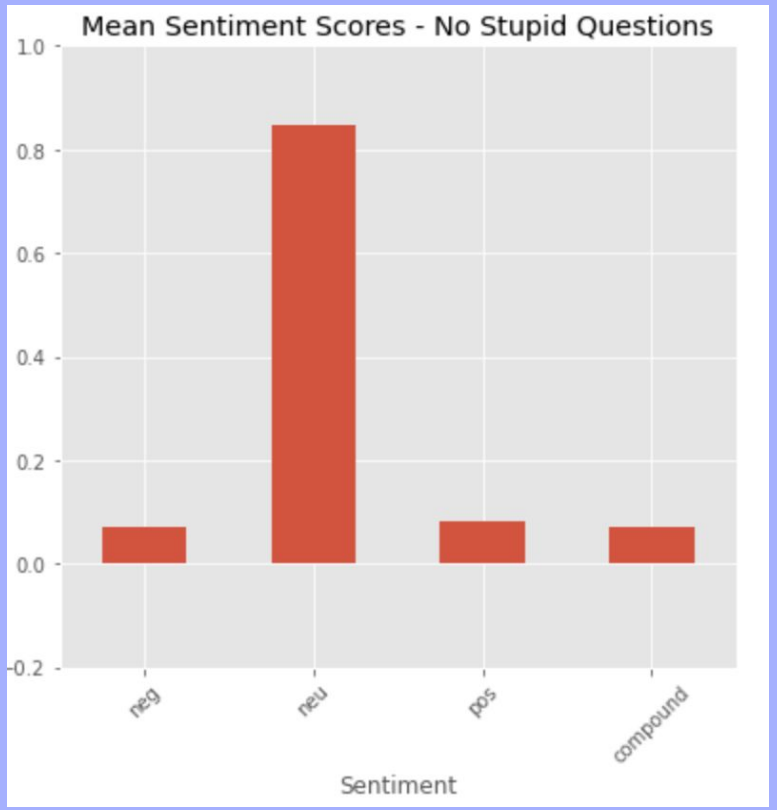
MOST FREQUENT WORDS





SENTIMENT ANALYSIS





Model 1:
NAIVE BAYES

GridSearchCV - best parameters

CountVectorizer

- Ngram_range: (1,2)
- Min_dif: 5
- Preprocessor: None
- Stopwords: english

Multinomial Naive Bayes

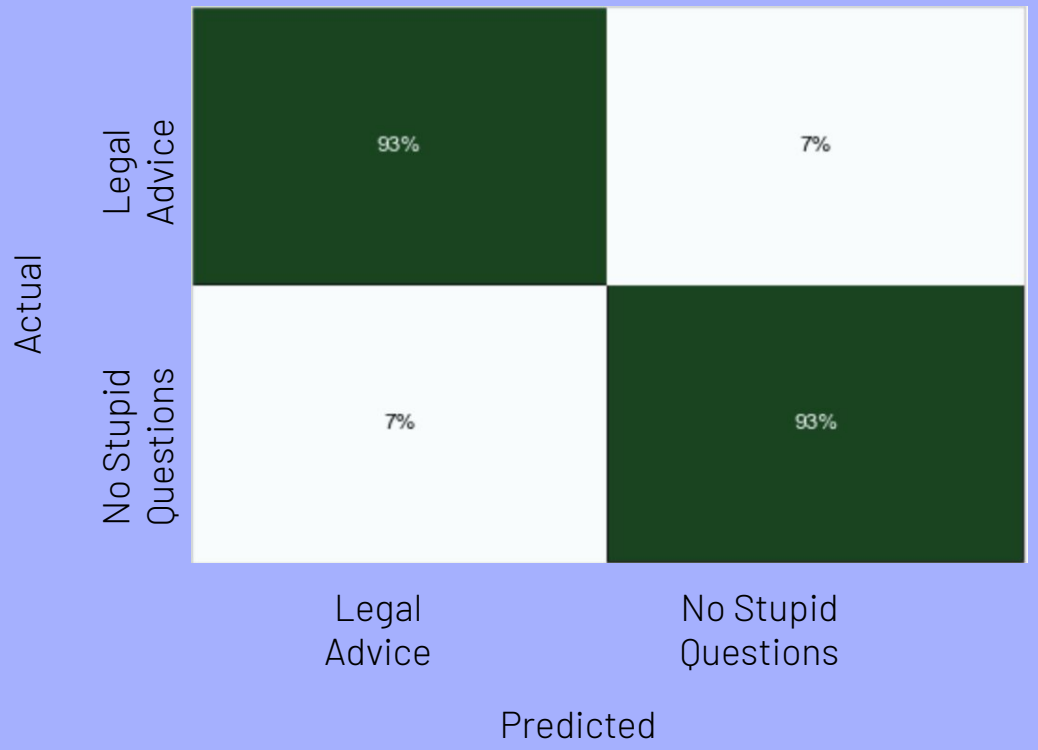
- Alpha = 1

Scores

- Train: .94
- Test: .93

MNB Model Performance

- F1 Score: 0.93
- Balanced accuracy score: 0.93

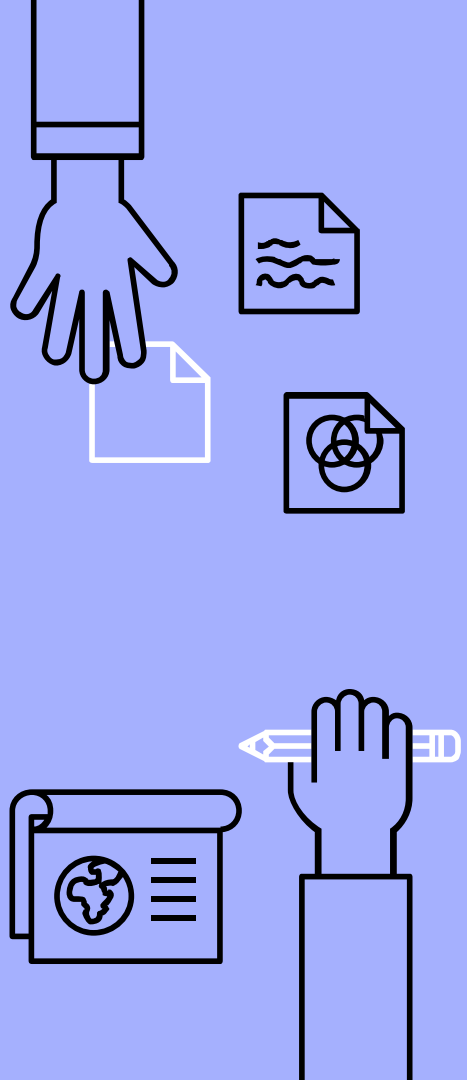


Model 2:

LOGISTIC REGRESSION

Model Exploration

	Training Score	Testing Score
Logistic Regression	0.992	0.949
K Neighbors	0.866	0.826
Decision Tree	0.999	0.868
Bagging	0.993	0.904
Random Forest	0.999	0.904
Adaboost	0.905	0.907
SVC	0.955	0.941



GridSearchCV - best parameters

CountVectorizer

- Ngram_range: (1,2)
- Min_df: 5
- Max features: 30,000
- Preprocessor: None
- Stopwords: English

Logistic Regression

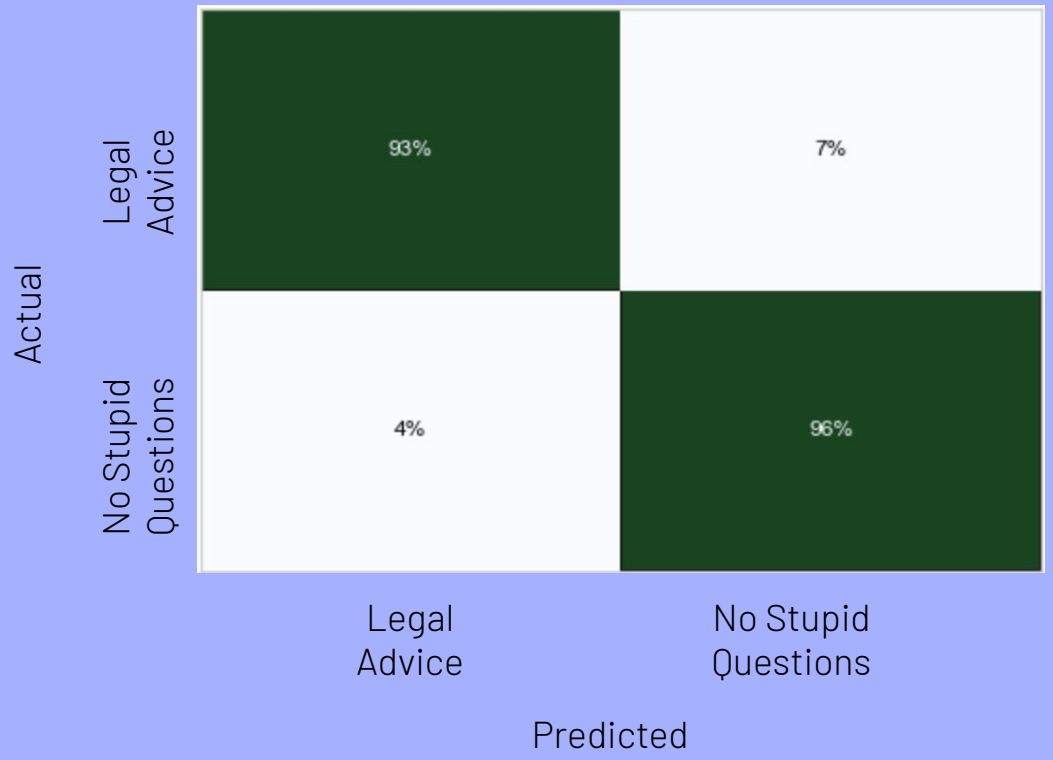
- C = 1.0

Scores

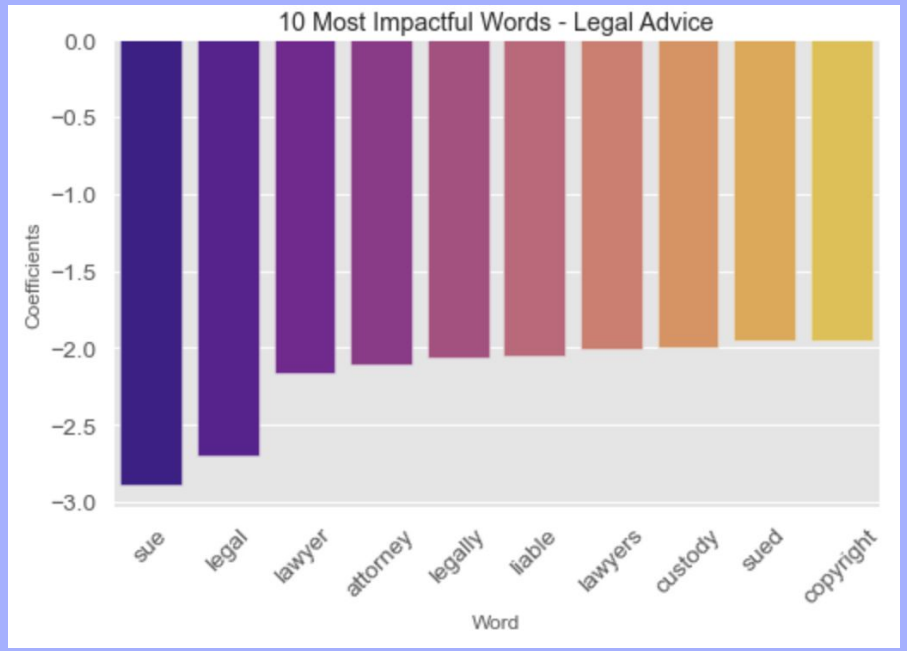
- Train: .99
- Test: .94

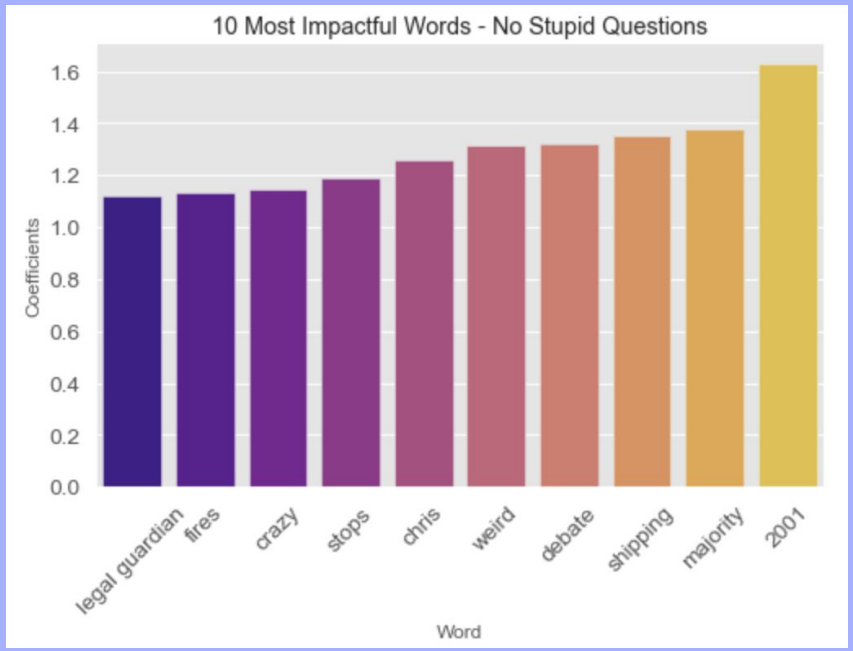
Logistic Regression Model Performance

- F1 Score: 0.934
- Balanced accuracy score: 0.934



Words with Impact

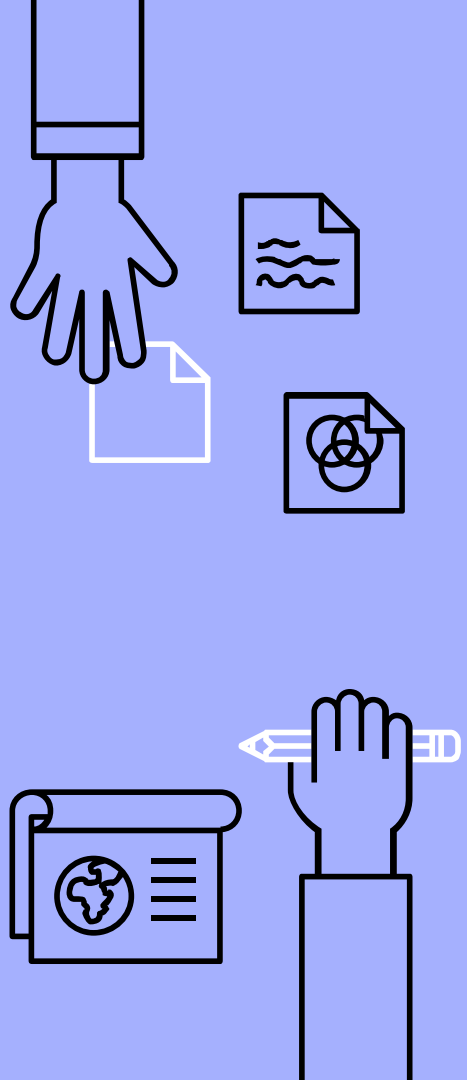




In Conclusion...

Findings

- Logistic Regression is top performer
- Interpretable coefficients
- Best test accuracy scores, balanced accuracy scores, and f1 scores
- Lawyers know where to spend their time



THANK YOU!