Helping Product Reviews be more Helpful

Predict Product Review Helpfulness using Machine Learning

Elliot Macy 01.18.21

Features

Textual and Contextual

Lexical

LEM/STEM: Normalized, stopped tokens

Syntactic

NOUN: Percentage of nouns

ADJ: Percentage of adjectives

ADV: Percentage of adverbs

VERB: Percentage of verbs

Structural

CHAR: Number of characters

NUM: Number of tokens

Structural (cont.)

WORD: Number of words

SENT: Number of sentences

INTERRO: Number of questions

EXCLAM: Number of exclamations

COUNT: Number of exclamation points

LEN: Average word length

AVG: Average sentence length

PER: Percentage of questions

CAPS: Percentage of capitalized words

Contextual

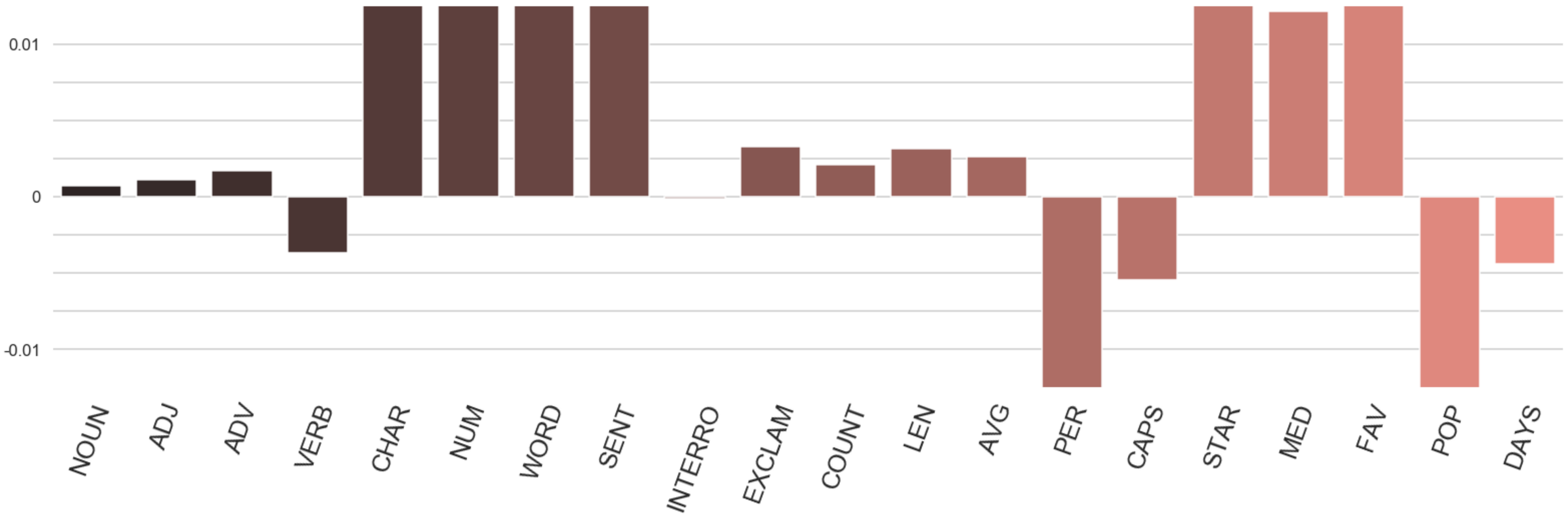
STAR: Reviewer’s star rating for product

MED: Product’s median star rating

FAV: Review’s star rating vs product’s median rating

POP: Number of product’s reviews

DAYS: Number of days from review date to product’s first review

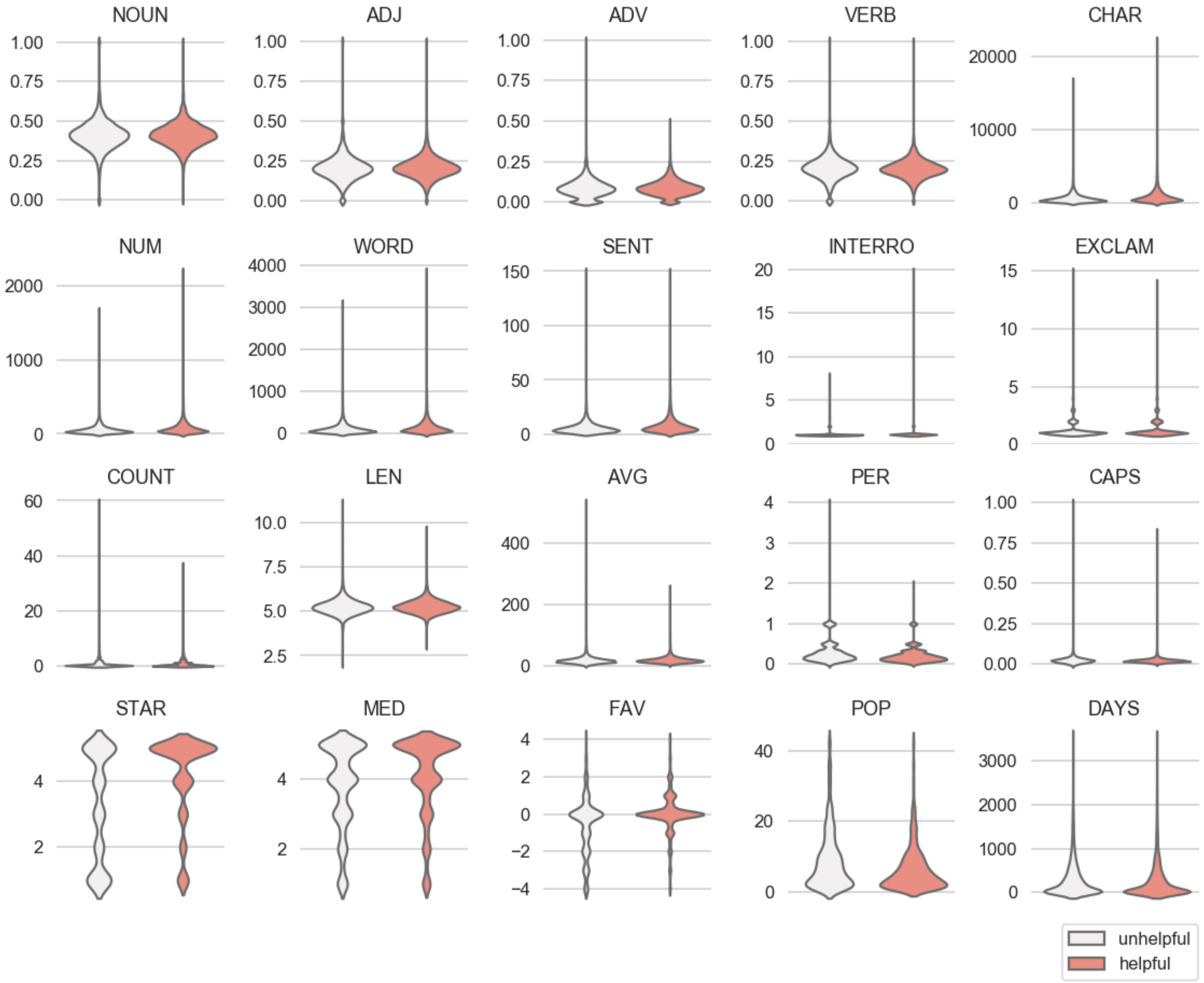


Correlation

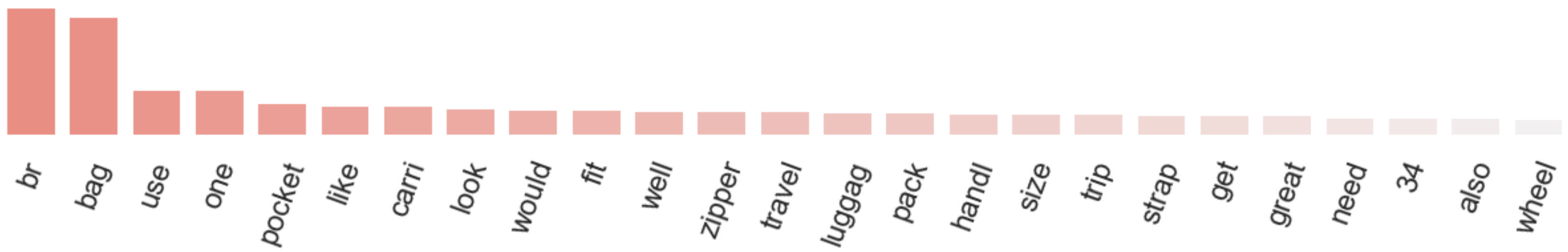
Feature-to-Helpfulness

Distributions

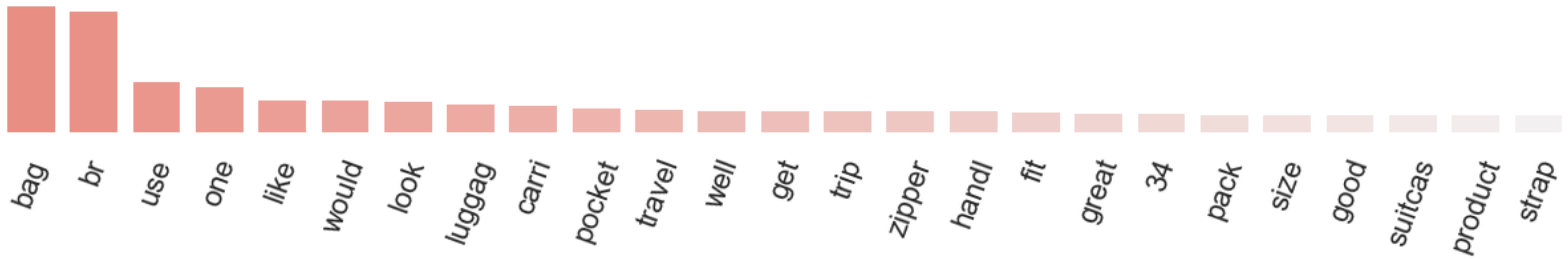
Helpful/Unhelpful



25 Most Frequent Helpful Review Terms

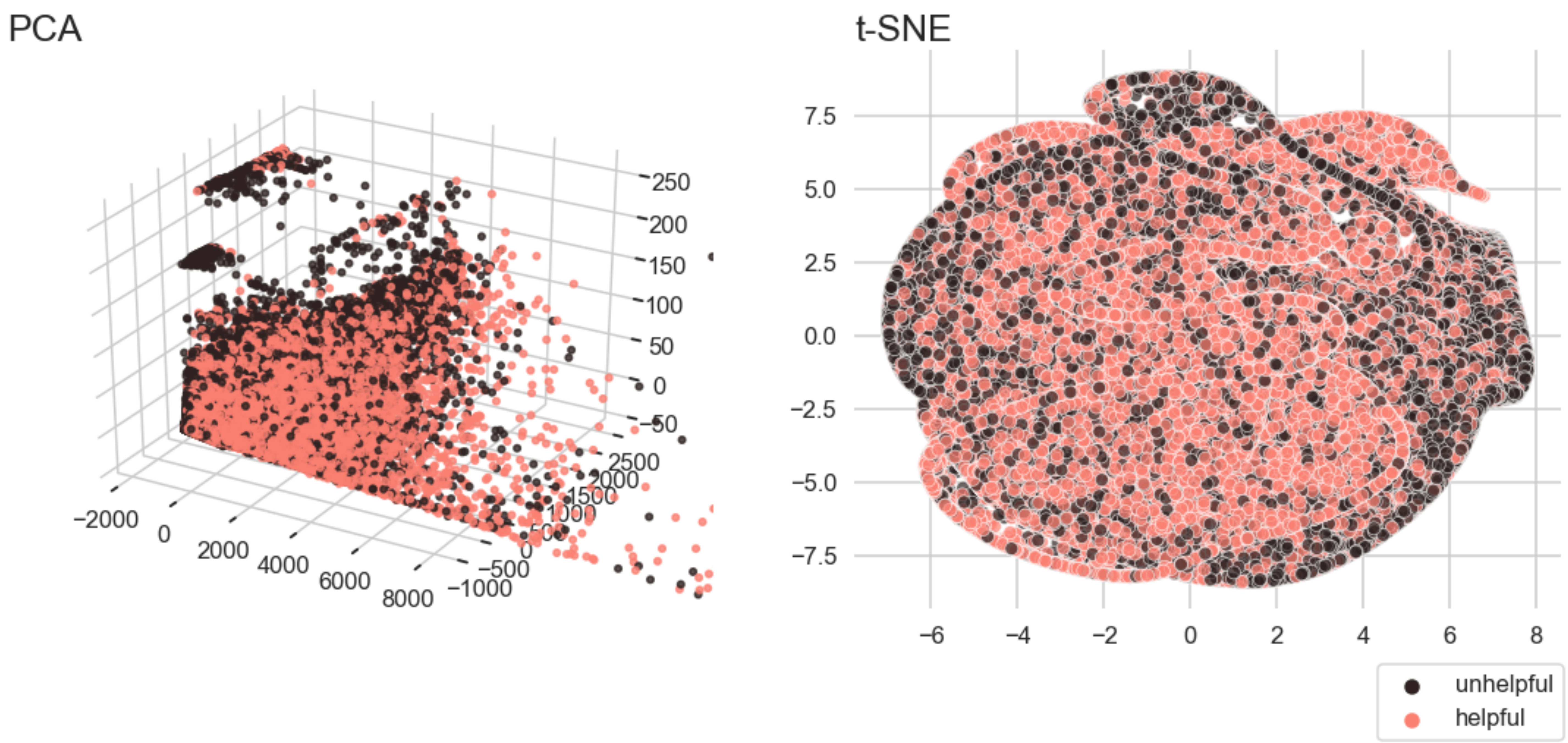


25 Most Frequent Unelpful Review Terms



Most Frequent Terms

Helpful/Unhelpful



Dimensionality Reduction

PCA and t-SNE

RANDOM FOREST

criterion = entropy, max depth = 2,
num estimators = 500,
max features = None

Training Accuracy:	0.6306
Testing Accuracy:	0.6294
Training Precision:	0.5933
Testing Precision:	0.5958
Training Recall:	0.7688
Testing Recall:	0.7714
Training F1:	0.6697
Testing F1:	0.6723

XGBOOST

min child weight = 2, subsample = 0.7,
learning rate = 0.1, max depth = 2,
num estimators = 300

Training Accuracy:	0.7048
Testing Accuracy:	0.6337
Training Precision:	0.6845
Testing Precision:	0.6168
Training Recall:	0.7309
Testing Recall:	0.6777
Training F1:	0.7070
Testing F1:	0.6458

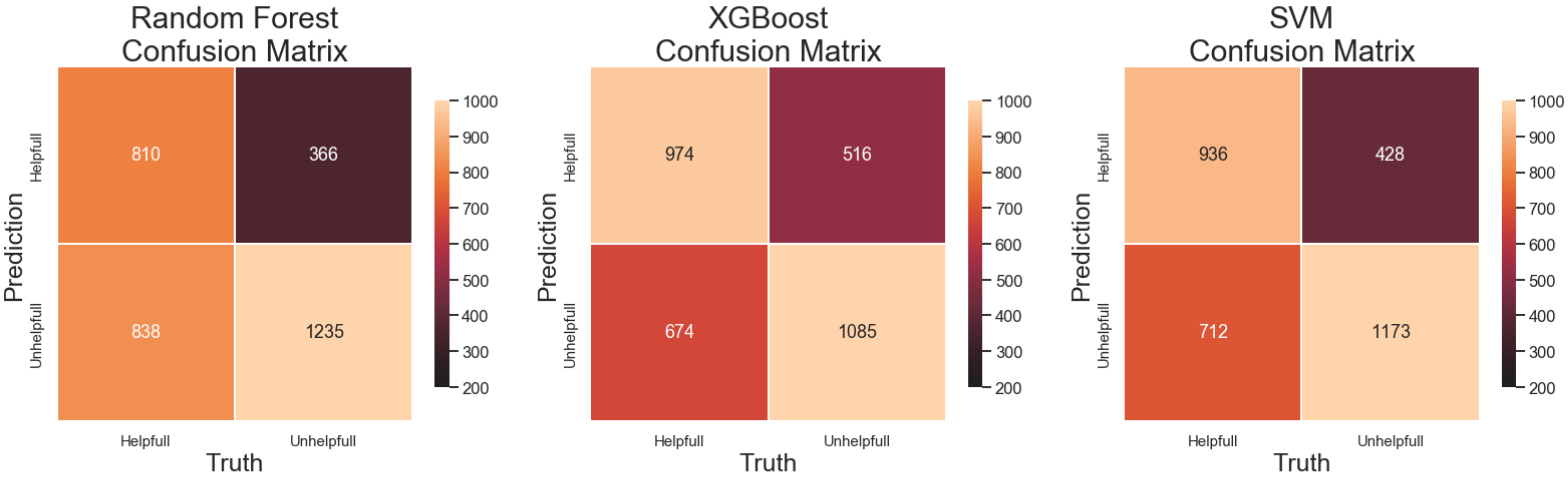
SVM

kernel = rbf, C = 2, gamma = 0.001

Training Accuracy:	0.6727
Testing Accuracy:	0.6491
Training Precision:	0.6389
Testing Precision:	0.6223
Training Recall:	0.7549
Testing Recall:	0.7327
Training F1:	0.6921
Testing F1:	0.6730

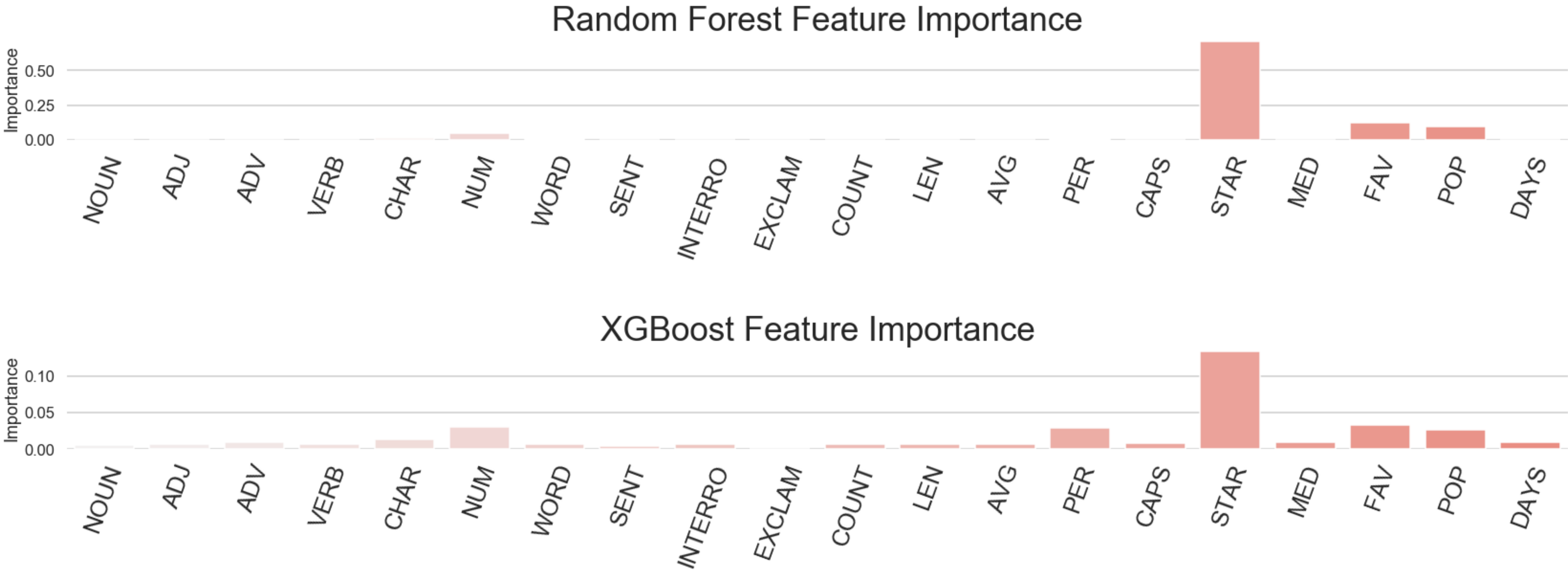
Results

Models, Hyperparameters, and Metrics



Results

Confusion Matrix



Random Forest Feature Importance

XGBoost Feature Importance

Feature Importance

Random Forest and XGBoost