Optimizing Access to Information by Differentiating Scientific and Simplified Questions on Reddit

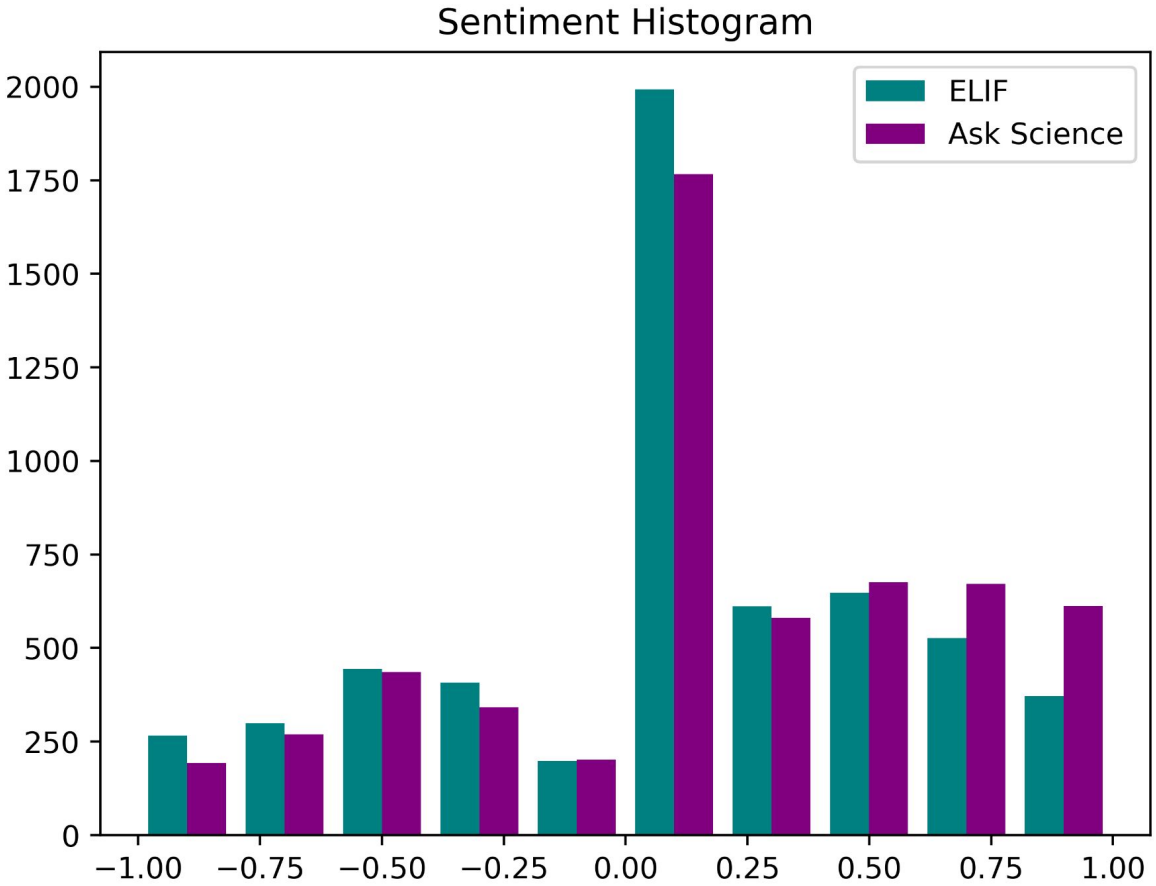
•••

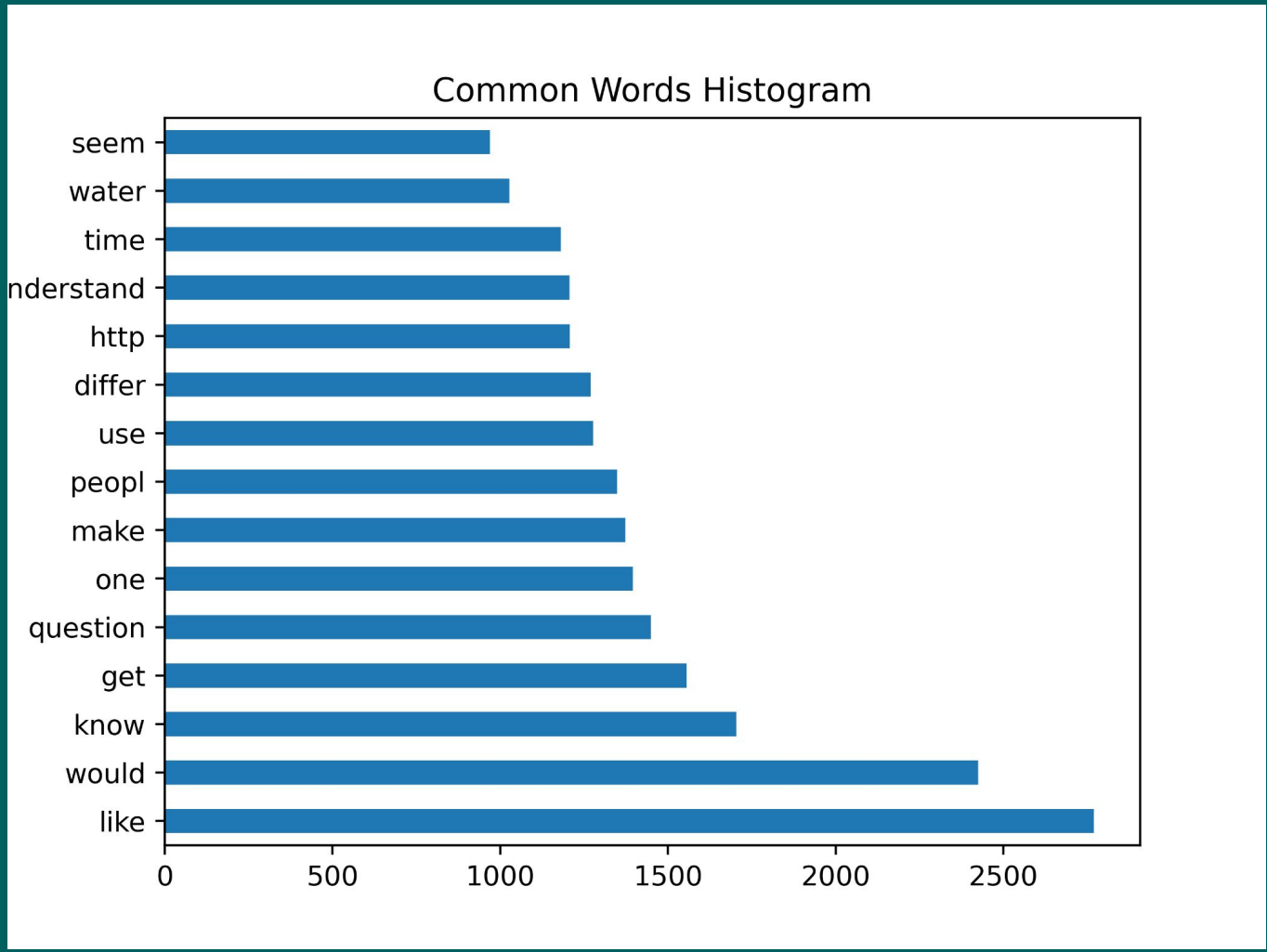
by Argishti Ovsepyan

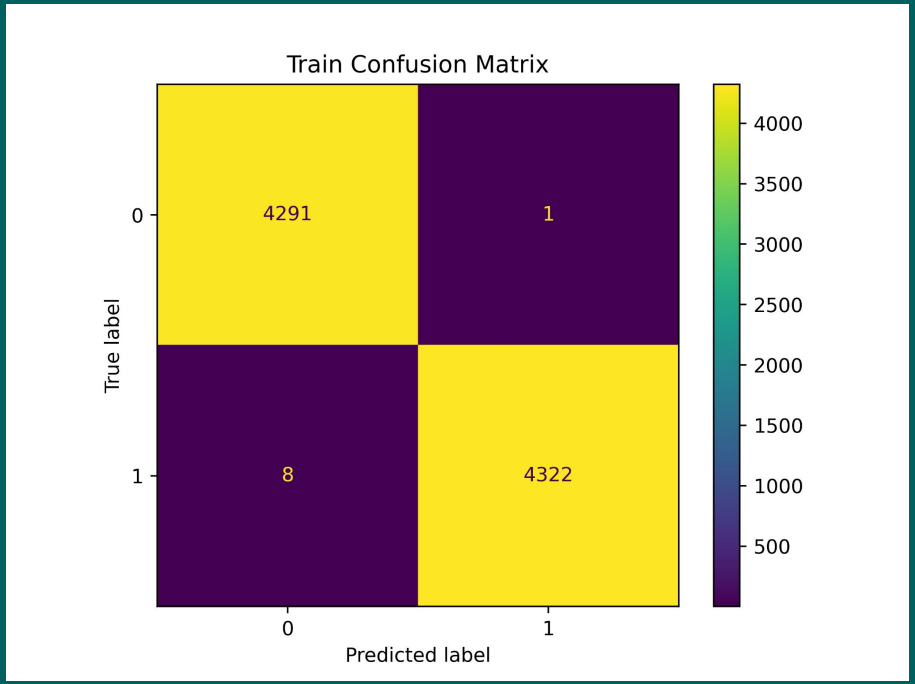
reddit

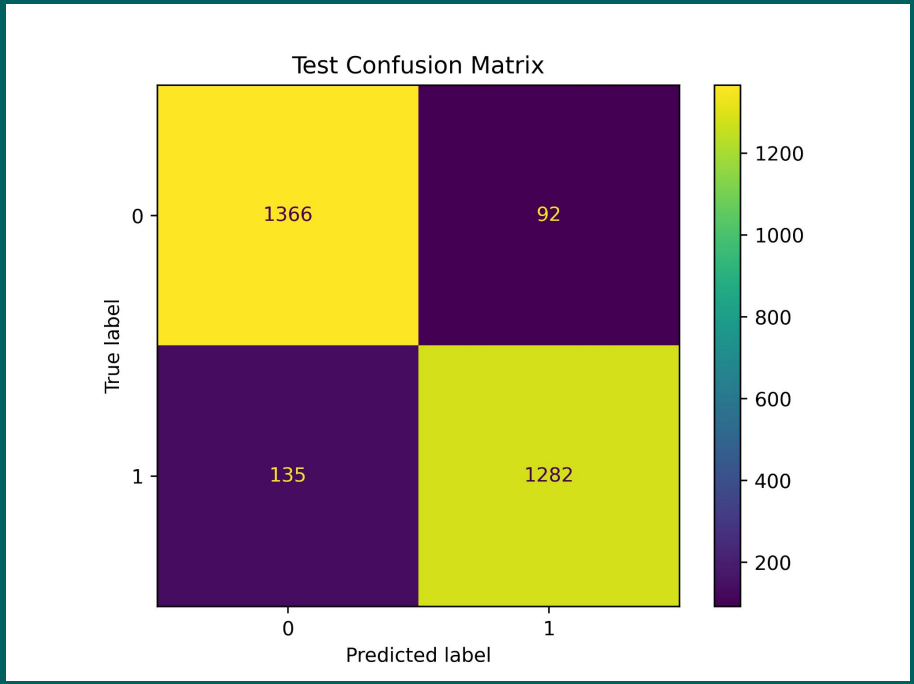
Problem Statement Outline:

- Classifying Reddit posts into two categories, complex from “Ask Science” and simplified from “explainlikeimfive” subreddits.
- Classification process is based on the specific language used within posts.
- The goal is to enhance user experience by facilitating access to information at different levels of detail depending on users input.
- Enable businesses to deliver personalized content and customer support tailored to each unique user.


Sentiment Histogram
ELIF
Ask Science
2000
1750
1500
1250
1000
750
500
250
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Common Words Histogram
seem
water
time
nderstand
http
differ
use
peopl
make
one
question
get
know
would
like
0
500
1000
1500
2000
2500





- Best model was logistic regression with ridge-like regularization and hyperparameter tuning.

- Best parameters were max features at none, min df as 1, max df as 0.4, ngram range (1,3), stop words at none, and a moderate regularization strength of 1.

- ROC AUC score of 0.998 on training data and 0.921 on testing data (original baseline of 0.50).

Key takeaways:

- **With a testing score of 0.921, the model demonstrates a strong ability to distinguish between complex and simple questions by effectively balancing the true positives and false positives.**
- **This is a valuable tool for classification between different types of inputs from unique users.**
- **Allows businesses to tailor responses for content delivery and customer support to the unique users level of understanding or appropriate level of explanation based on their input level of complexity versus simplicity.**

Questions?