Classifying Showerthoughts and casualConversation

By Daniel Kim

Situational Context

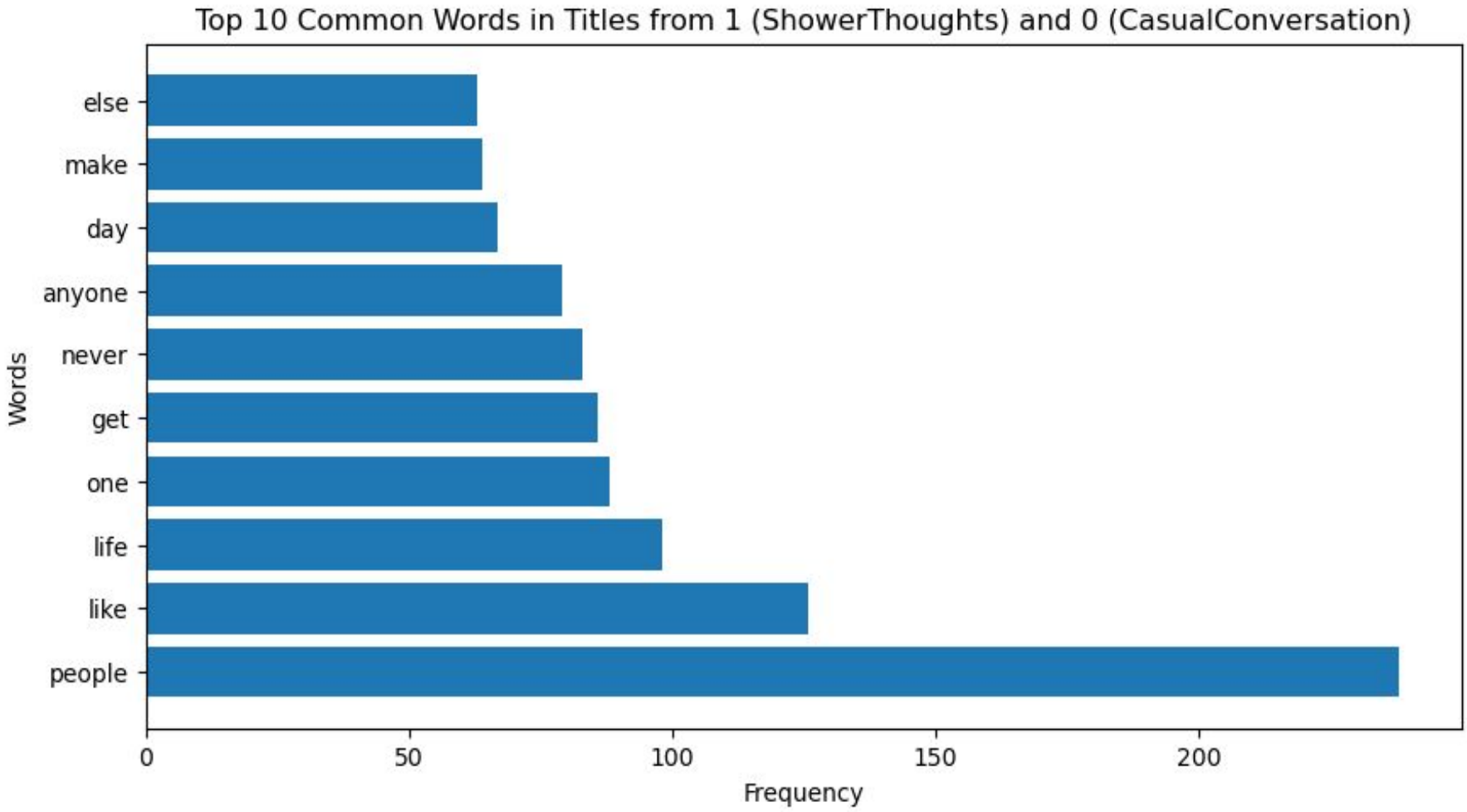
You are hired by a social media startup that aims to create a personality quiz app that can classify users as either happy or sad based on their responses to selected questions. Your task is to help the team build the app’s backbone by creating a model and methodology that the organization's team can replicate and adjust to fit their app's specific needs.

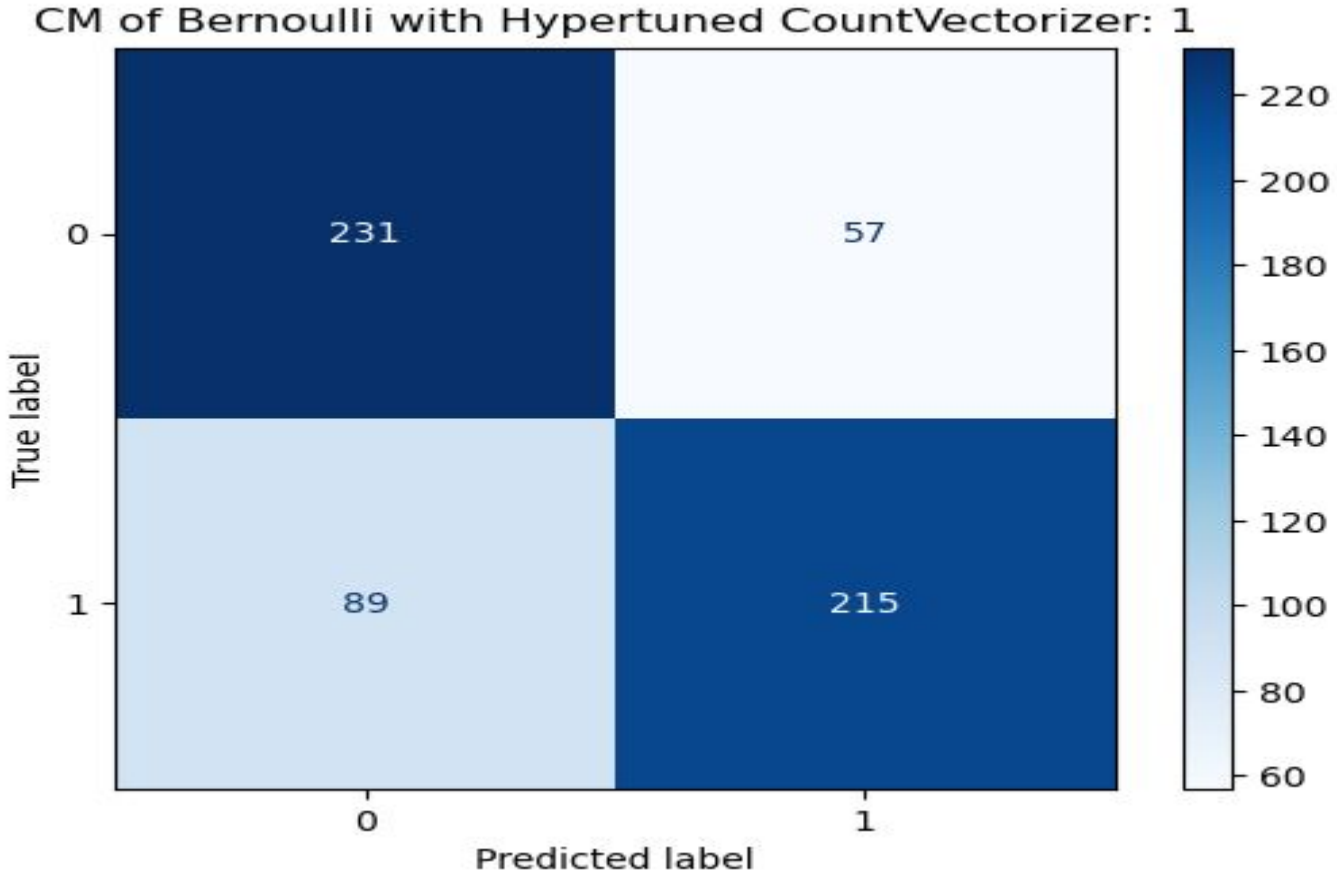
Problem Statement

To help the social media startup's team establish the backbone of their personality quiz app, this project aims to collect data from two subreddit communities: Showerthoughts and CasualConversation. Leveraging the subreddit data, this project aims to train the data to create and compare four different models' performance accuracy rates using the ROC-AUC and test scores as success metrics. The project will recommend that one model and its methodology based on the highest performing ROC-AUC and test scores.

Word count: 4973

Average Frequency: 2.92


Top 10 Common Words in Titles from 1 (ShowerThoughts) and 0 (CasualConversation)
Words
else
make
day
anyone
never
get
one
life
like
people
0
50
100
150
200
Frequency


CM of Bernoulli with Hypertuned CountVectorizer: 1
231
57
89
215
0
1
True label
0
1
Predicted label
220
200
180
160
140
120
100
80
60

Recommendation based on ROC and Test Scores

Summary of Test and ROC Scores for Each Model for second iteration:

- CV stands for CountVectorizer
- TV stands for TfidVectorizer

Model	CV Test Score	CV ROC Score	TV Test Score	TV ROC Score
Logistic Regression	0.785472972972973	0.785453216374269	0.7820945945945946	0.7811586257309941
Bernoulli	0.7432432432432432	0.744517543859649	0.7432432432432432	0.744517543859649
Multinomial	0.6942567567567568	0.6911549707602339	0.6976351351351351	0.6937134502923976
Random Forest	0.7820945945945946	0.7813413742690059	0.7736486486486487	0.7723866959064327

Summary of Test and ROC Scores for each Model for first iteration:

Model	CV Test Score	CV ROC Score	TV Test Score	TV ROC Score
Logistic Regression	0.7905405405405406	0.7902046783625731	0.7820945945945946	0.7808845029239766
Bernoulli	0.7533783783783784	0.7546600877192983	0.7533783783783784	0.7546600877192983
Multinomial	0.7010135135135135	0.6977339181286549	0.6959459459459459	0.6921600877192982
Random Forest	0.785472972972973	0.7840826023391813	0.7804054054054054	0.7792397660818713

Conclusion and next steps

Here are following next steps to continue to improve the recommended model and investigate other findings from the four models:

- Experiment with Logistic Regression's hyperparameters such as adjusting C to reduce model's overfitting
- Experiment with other models' respective hyperparameters such as adjusting alpha and minimum sample leafs to reduce overfitting
- Investigate why overall, the test and ROC scores are generally similiar to one another
- Investigate why the Bernoulli model is the only one that was better at identifying CasualConversation than Showerthoughts
- Investigate why Logistic Regression's test scores with TfidVectorizer from both iterations remain the same
- Collect more unique subreddit posts to increase the diversity of words that our models can train on

References

- Happy/Sad Emojis
- Getty Images/iStockphoto
- Showerthoughts Logo
- https://www.reddit.com/r/showerthoughts
- CasualConversation Logo
- https://www.reddit.com/r/CasualConversation/