Evaluating the Differences between r/lifehacks and r/LifeProTips

Franco Sanchez

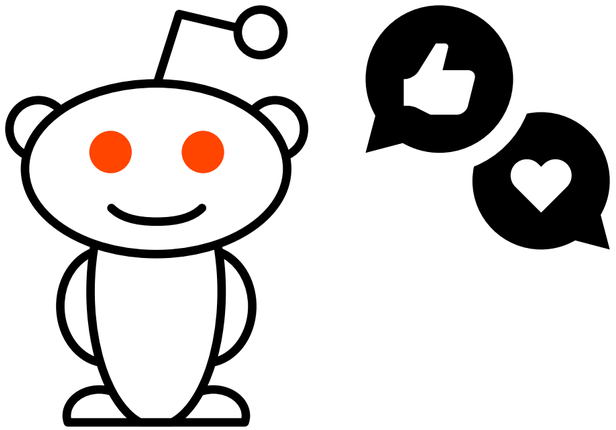
Problem Statements

- Create a model that predicts if a Reddit post belongs to r/lifehacks or r/LifeProTips, based on post contents.

- Identify the top 10 keywords that distinguish both subreddits.

Background

Two subreddits dedicated to sharing advice that makes life easier.



r/lifehacks (LH)

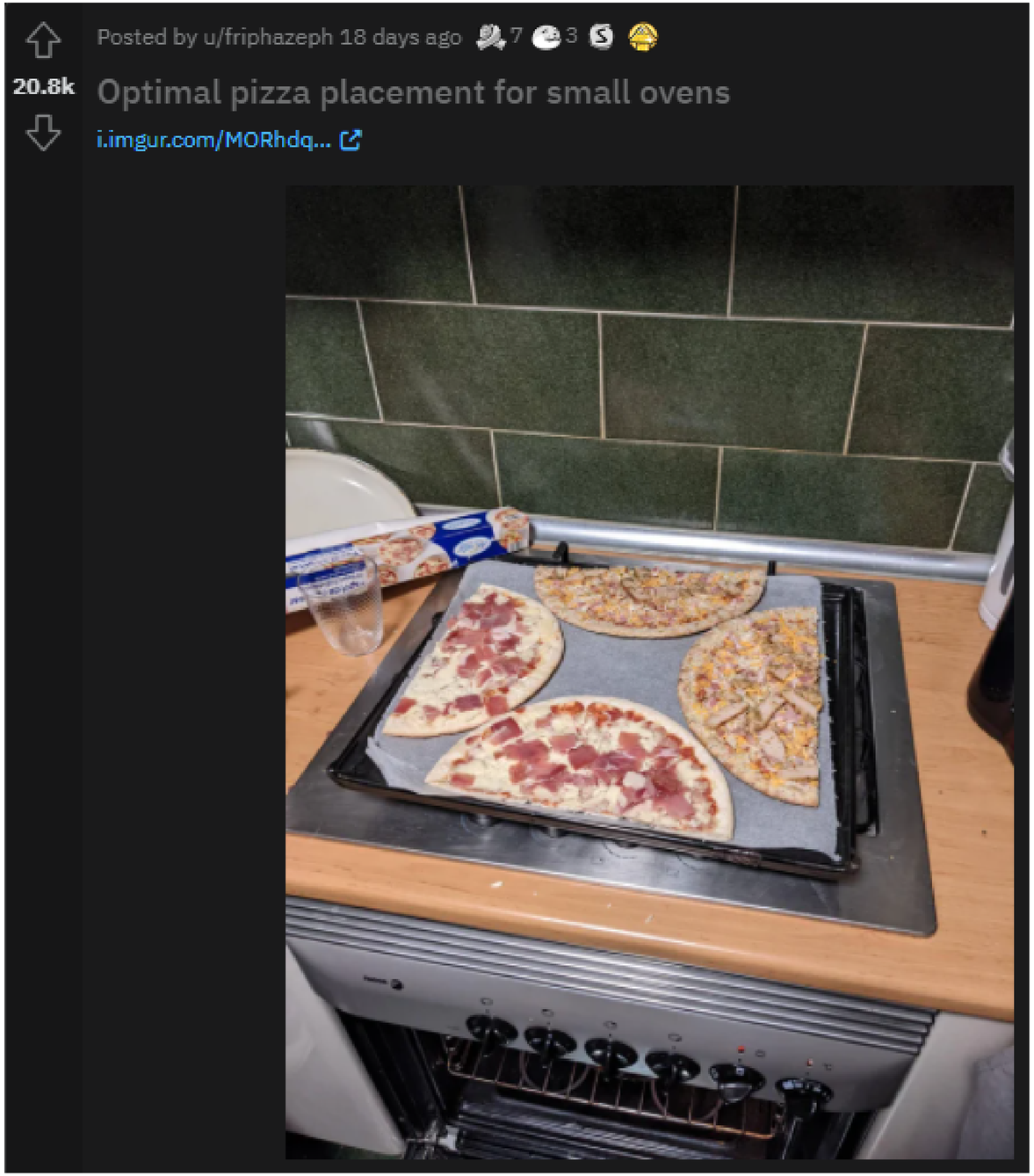
"Lifehacks: Uncommon solutions to common problems."

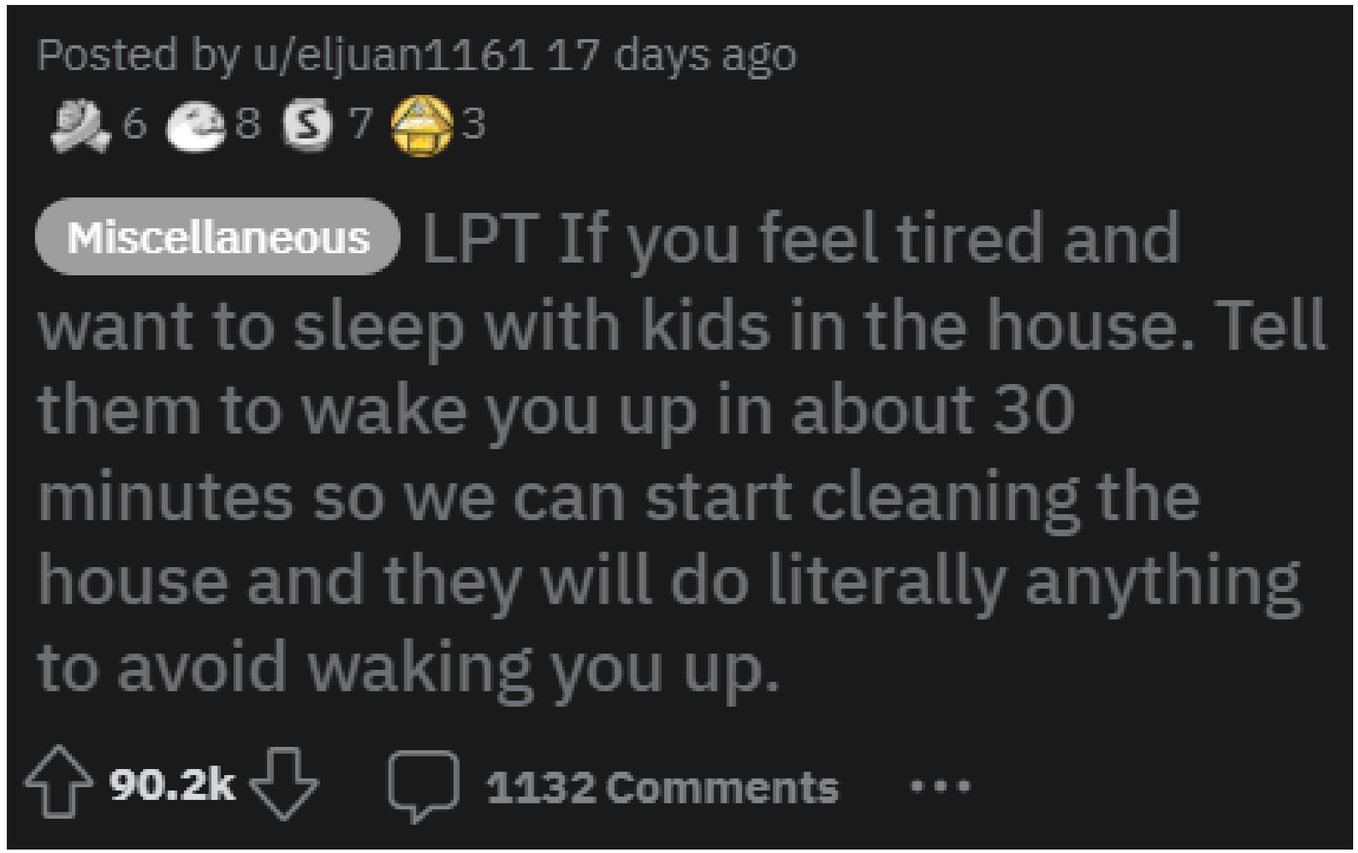
r/LifeProTips (LPT)

"Tips that improve your life in one way or another."

Example

r/lifehacks top post from April.





Example

r/LifeProTips top post from April.

Modeling Process

01
Data wrangling

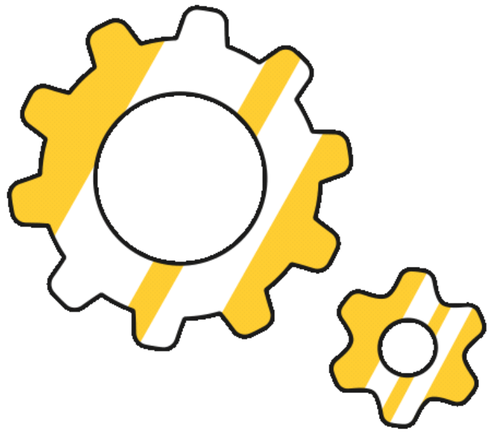
- Gathered data using the Pushshift Reddit API
- Aggregated data into single data frame

02
Data preprocessing

- Dropped nulls and other erroneous data
- Aggregated into single clean data frame
- X and y train-test splits

03 - 04
Feature Engineering

- Vectorization
- Tokenization
- Lemmatization



03 - 04
Modeling

- Classification models
- Tuning hyperparameters
- Performance

05
Insights

- Classification metrics
- Inferences

The Data

Feature Names	Descriptions
selftext	Body text of post
subreddit	r/lifehacks (1) or r/LifeProTips (0)

Rows: 2,476

Columns: 2

Training split: 70%
Test split: 30%

LH: 0.241

LPT: 0.759

**imbalanced data proportions*

The Models

Tried various models:

- Logistic Regression
- KNN
- Bagged Trees
- Random Forests
- Extra Trees
- Ada Boosting
- Ensembles

Baseline

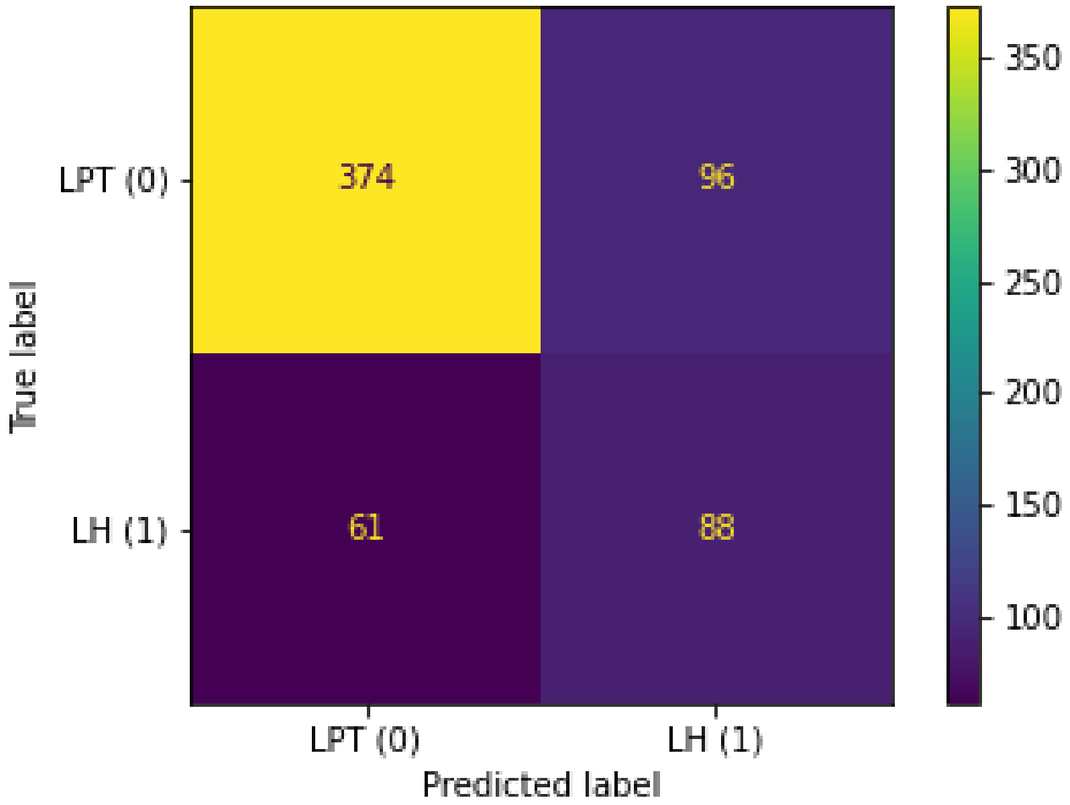
Accuracy	0.759
Recall	0
Specificity	1
Precision	Null (0)
F1 Score	0

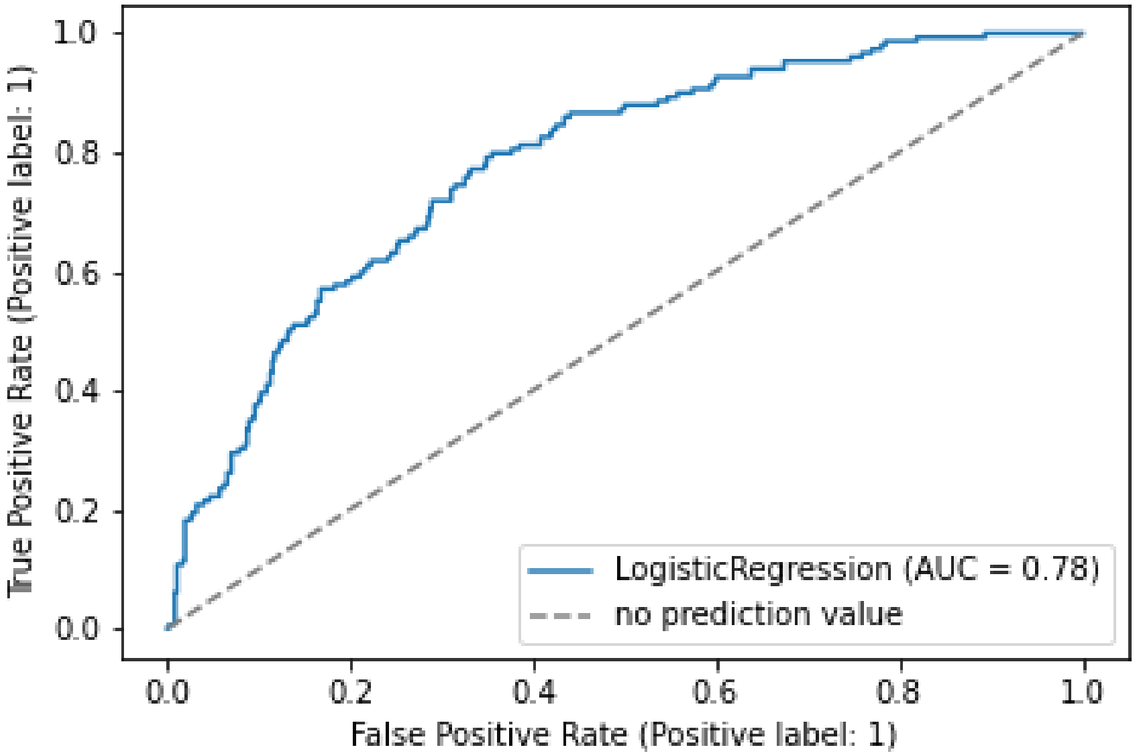
Top Scores

Accuracy	0.746
Balanced Accuracy	0.693
Recall	0.591
Specificity	0.796
Precision	0.478
F1 Score	0.528

Best performing: **Logistics Regression**

Performances

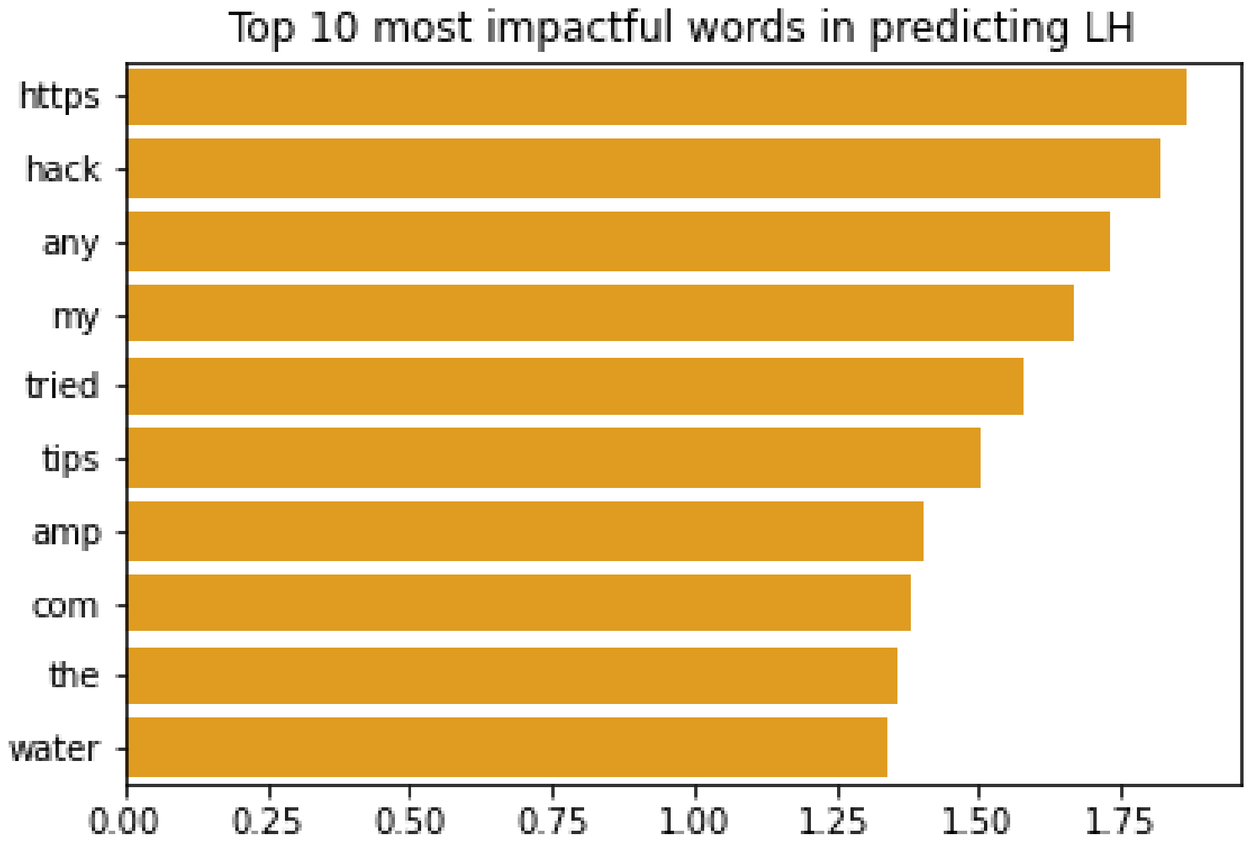




Insights

Classes

Label	Count	Proportion
True Positive	88	0.48
False Positive	96	0.52
False Negative	61	0.14
True Negative	374	0.86



**TF-IDF tokenization*

Metrics Inference

Recall	The proportion of correctly predicted r/lifehacks posts over actual r/lifehacks posts
Precision	The proportion of correctly predicted r/lifehacks posts over all r/lifehacks predictions
F1 Score	Harmonic mean of recall and precision

Discussion

Imbalanced data requires more feature engineering.

- Add more relevant stopwords
- Oversampling / undersampling
- Eliminate more erroneous posts (bots/spam posts)

Shift classification threshold to account for imbalanced data.

Wrangle more balanced data from the subreddits.

Thank you