Understanding the Customer Voice Through NLP

Brian Cho, General Assembly DSIR 2-8

Agenda

1. Project Concept
2. Procedure
3. Initial Analysis
4. Evaluating the Model
5. Implementation

Project Concept

Business Core Value: Adopting a customer-centric strategy

Create a machine learning model to feed Lyft with customer feedback in order to drive real-time improvements

**Companies who use a customer-centric strategy are 60% more profitable [source: Deloitte]*

Customer-Centricity: Embedding it into your organization's DNA

Procedure

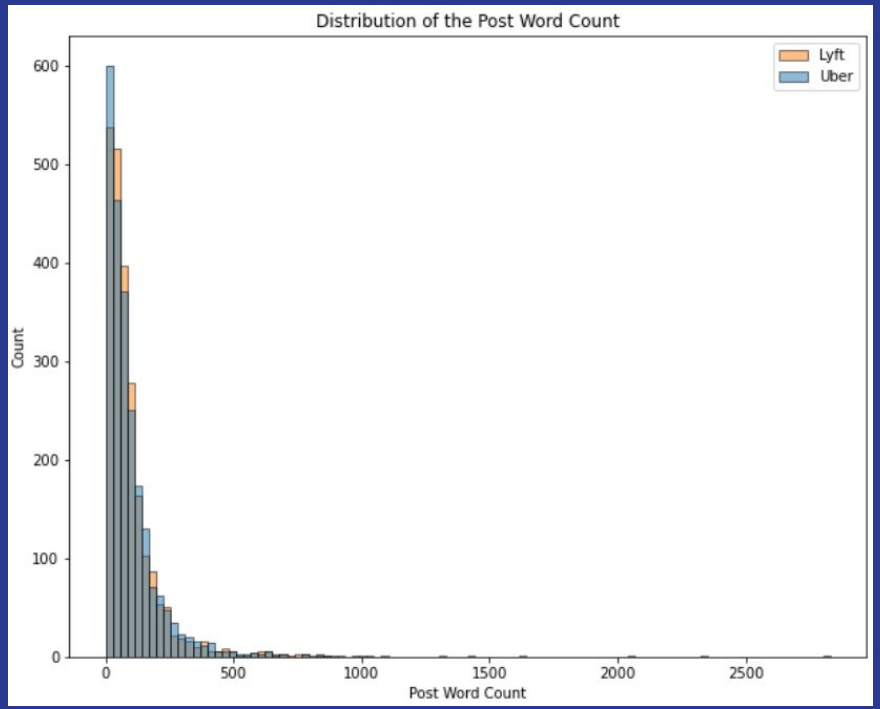
** Example model will identify text data and classify whether it came from the Lyft or Uber Subreddit*

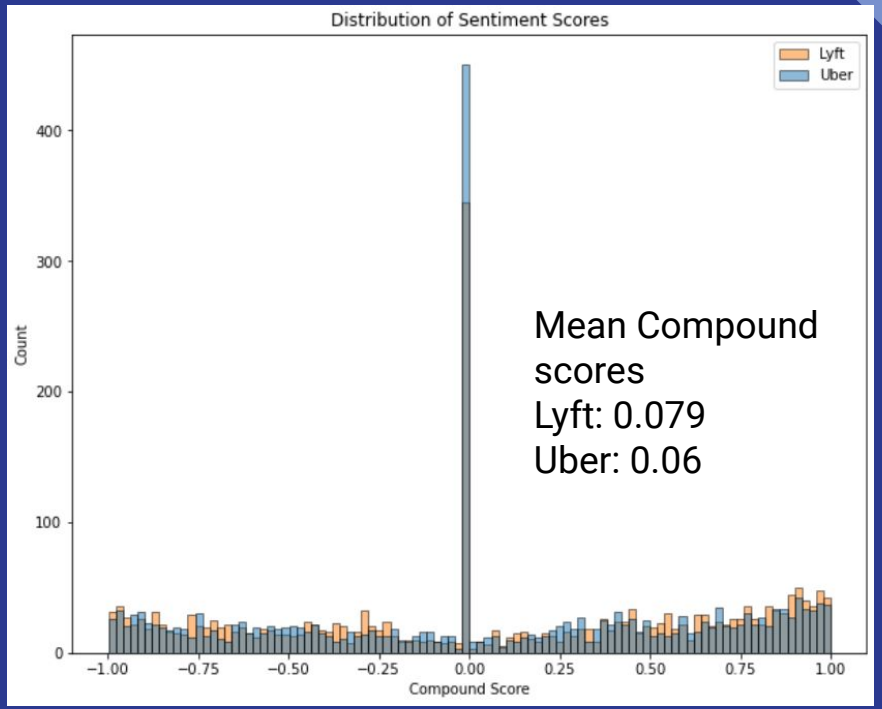
Data Gathering	Initial Analysis	Modeling Data	Evaluate	Refine Model
~7000 Reddit Posts - Large Sample Size -Equal Distribution of Uber and Lyft posts	-Clean Data -Filter Stop Words -Top Words -Does the Data Make Sense? -Sentiment Analysis	-Transform Data -Model Baseline -3 Classification Models: Naive Bayes, Logistic Regression, Random Forest	-How does it compare to the baseline model? -Success Metrics	Explore methods to refine model

Data Considerations

1. Stop Words
2. Null Values
3. Outliers
4. Sentiment Analysis

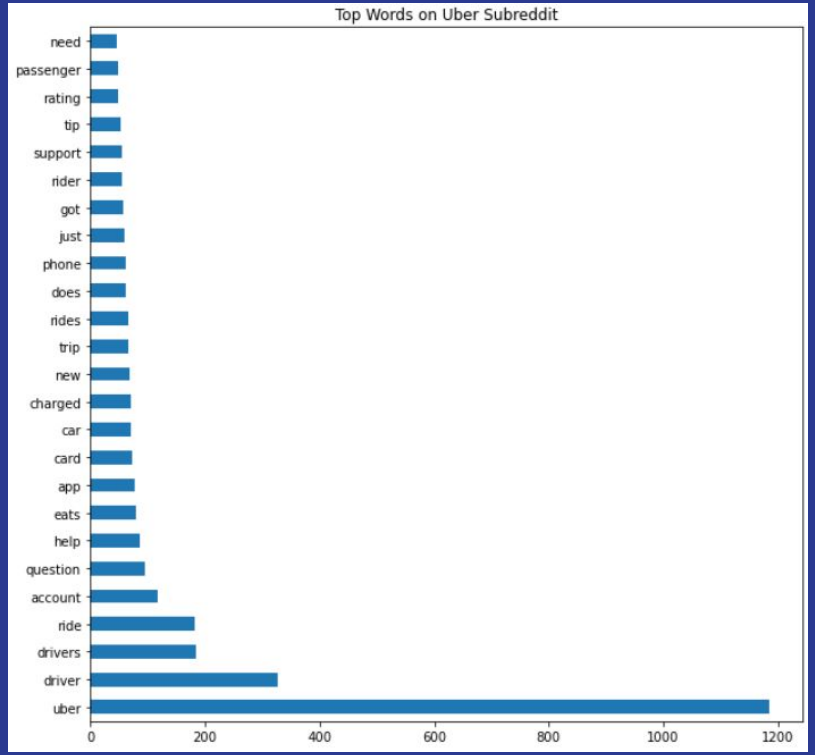
Understanding the Data

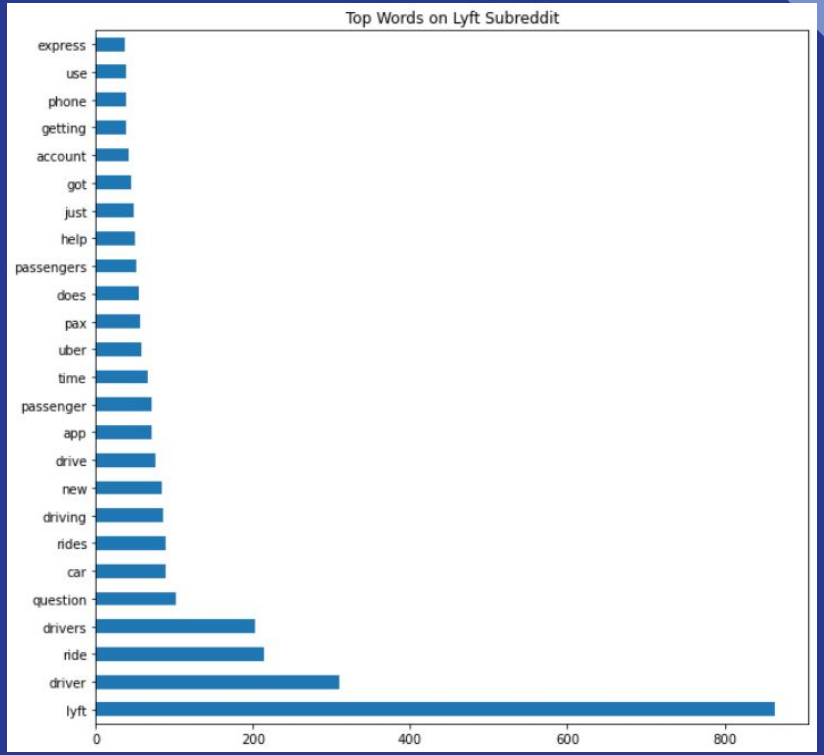




Mean Compound scores
Lyft: 0.079
Uber: 0.06

Understanding the Data





Evaluating the Model

Baseline Accuracy: 50% (n=4663)

Metrics	Naive Bayes	Logistic Regression	Random Forest
Accuracy	0.79	0.84	0.83
Precision	0.80	0.85	0.84
Recall	0.80	0.84	0.83
Train Score	0.86	0.90	0.83
Test Score	0.80	0.84	0.82

Looking Ahead

Compared to the baseline accuracy of 50%, the logistic regression model predicts with 84% accuracy.

Serves as a foundation to pull the customer voice on other social media platforms: Twitter, Google Reviews, Yelp, Facebook, etc.

Provides the company with real time feedback from customers

Sentiment Analysis scores can serve as a company wide metric to improve upon

Drive better strategic decisions