Data Set: A Musical Instruments Reviews from Amazon

Data Brief

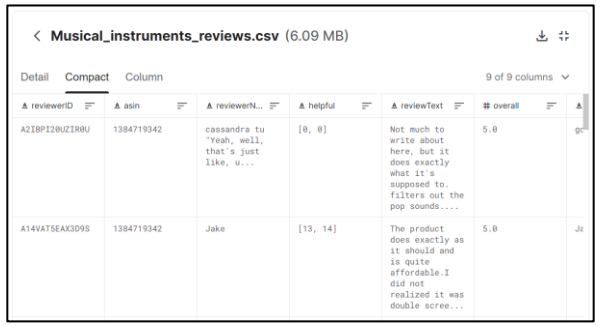
The Data Set

- Comes from kaggle
- Collected 2 years ago
- Has 10,261 unique reviews from 1,429 viewers
- Includes data fields of
  1) helpfulness on rating,
  2) text of the review,
  3) rating of the product,
  4) summary of the review,
  5) product ID
  ……

Possible Shortcoming

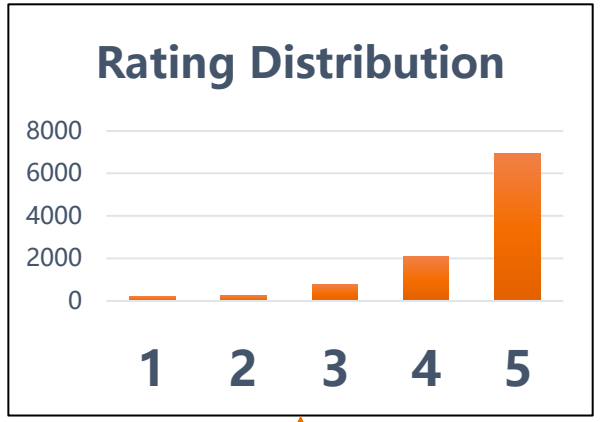
1) Approximate 10,000 samples maybe too small to have sufficient statical power

Solution: Run a k-fold validation to improve the performance



2) The distribution of product rating score is biased.

Solution: Including the helpfulness rating score can compensate to this issue.



Data Visualization Method

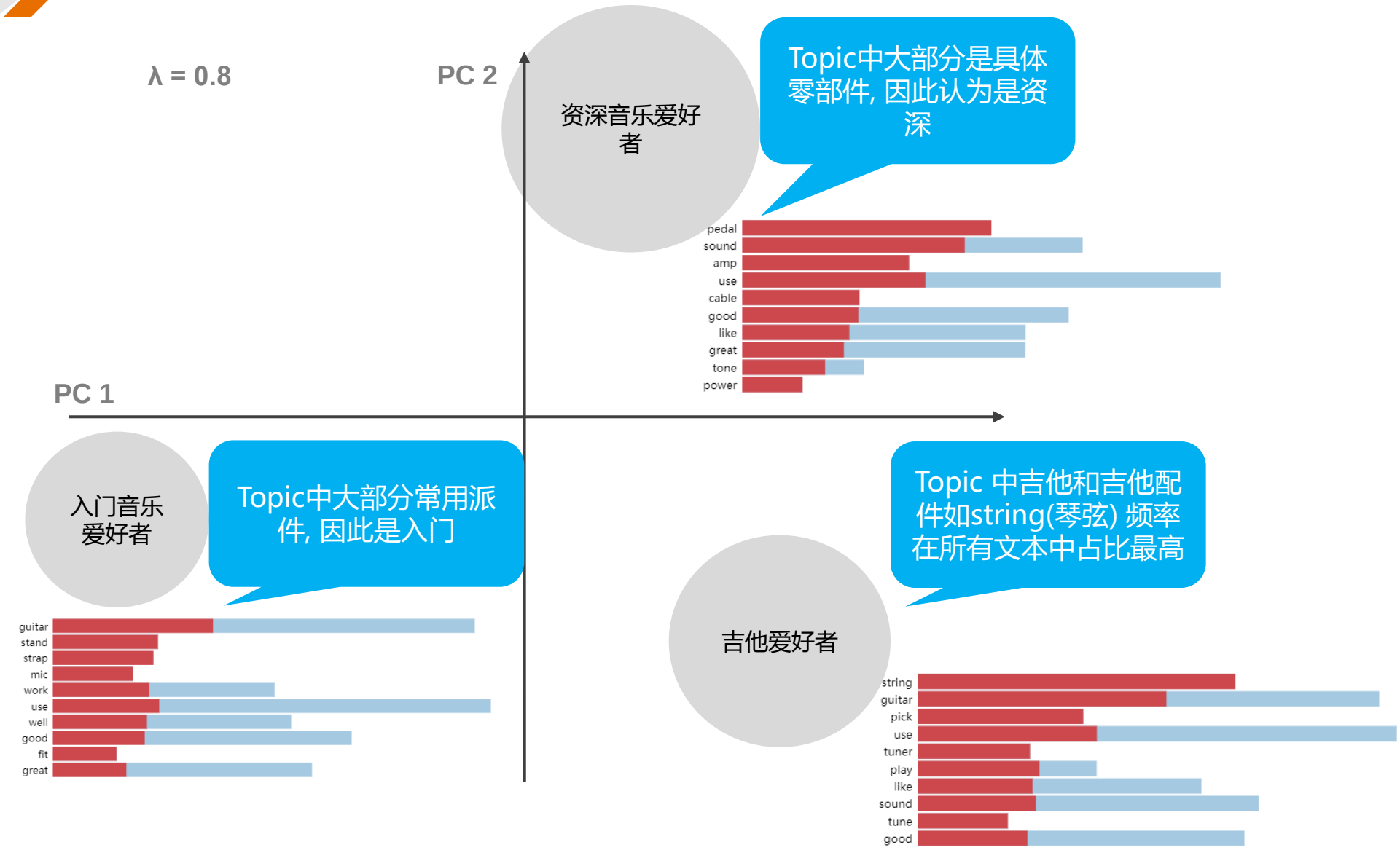
Method: Word Cloud

Method: Histogram

Data Set Source: https://www.kaggle.com/eswarchandt/amazon-music-reviews?select=Musical_instruments_reviews.csv

Topic Modeling: Unsupervised and Supervised Methods

Draft



Data Preprocess: Up Sampling + Down Sampling to balance data distribution

Draft

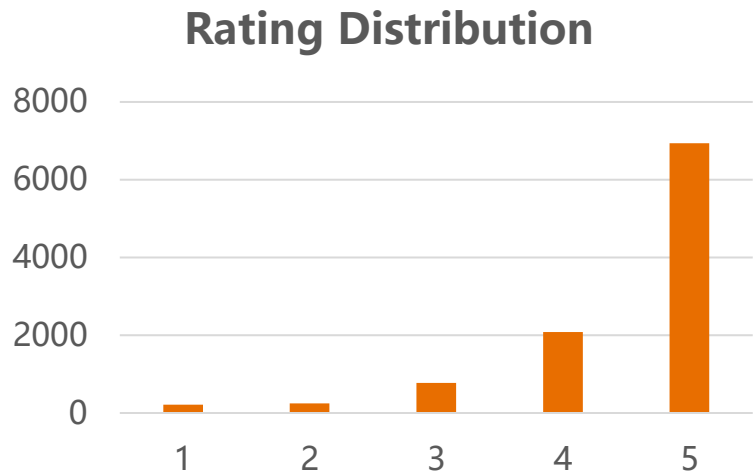
上采用: 利用回译增加偏少的class的样本量

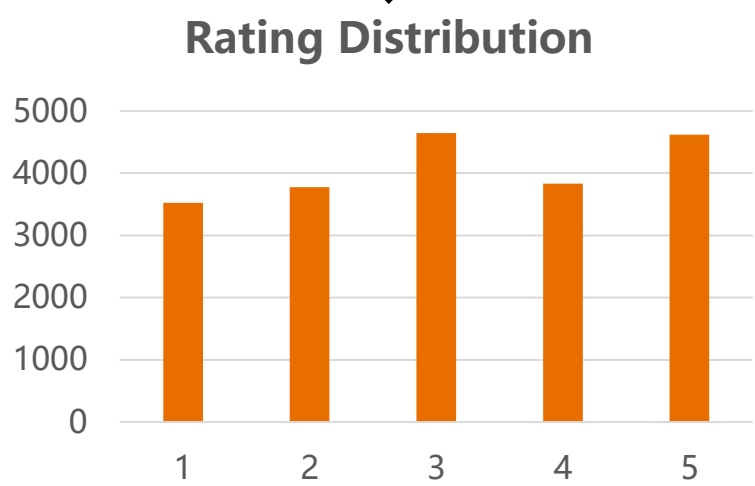
Up Sampling

Text Augmentation – Back Translation to generation similar sentence for minority class

- **Original Sentence:**
- The improvement over the old formulation is noticeable.

- **Sentence After Back Translation:**
- The improvement is obvious compared with the old formula.
- This is a significant improvement over the previous recipes.
- Improvements in old formulations are remarkable.





Down Sampling

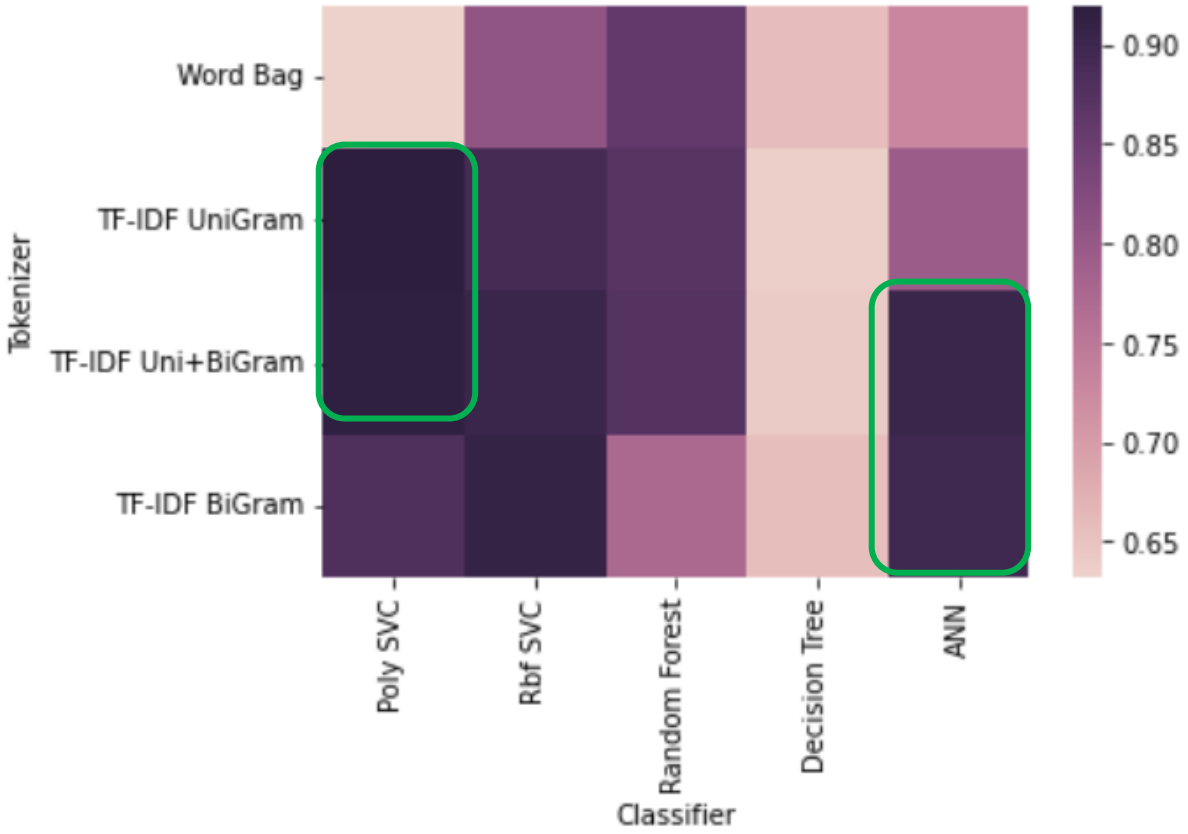
Delete random sample according to total distribution of sample class for majority class

下采用: 将偏多的class的样本随机删除, 使总分布平衡

Topic Modeling: Unsupervised and Supervised Methods

Draft

Model Performance on Test Set



Insight

- Compared to word bag, TF-IDF is a better way to token in a comment environment.
- ANN can do better on high dimension token Uni+Bi Gram TF-IDF , but will overfit on simple word bag or Uni Gram TF-IDF
- Poly SVC do good in both high and low dimension token.
- ★Based on the model, manager do not need to label the rating manually in a non-survey environment like comment under Youtube video regarding his product.

Objective: A model provide two kinds of scores from text comments on products

Objective

Provide a model that can estimate the customer's product rating score and helpfulness score based on the text comments on the product

Text Comments

Product Rating Score (5.0/5.0)

Comment Helpfulness Score (2/3)

Application

1) Comment Value Evaluation

Target E-commerce website (Amazon, Ebay, etc.)

Prioritize popular comments by making adjustment in online display and recommendation, which helps customers better understand the product then increase sales

2) Sentiment Prediction

Target video platform (YouTube, TikTok, etc.)

Generate sentiment prediction based on comments in a non-survey environment to have a better command on product property.

3) Establish label data storage

Target analytic companies

The model can generate text labels for no-rating reviews for further use.

Data Visualization Method

Method: Word Cloud

Method: Histogram

Sentiment Analysis: Unsupervised and Supervised Methods

VADER Lexicon-Based (Unsupervised)

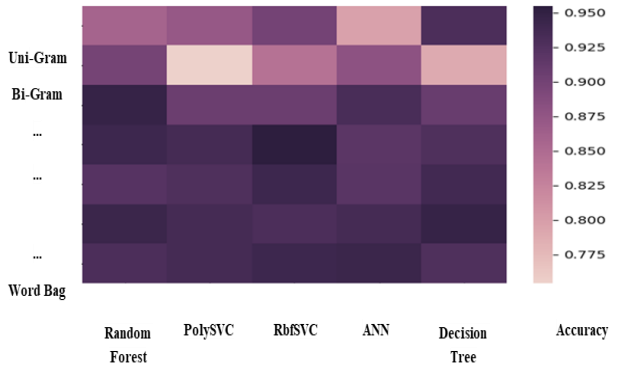
Use Valence Aware Dictionary and sEntiment Reasoner (VADER) lexicon to tell the polarity (positive /negative) and intensity

Polarity

Intensity

PRO: No human-created labels needed

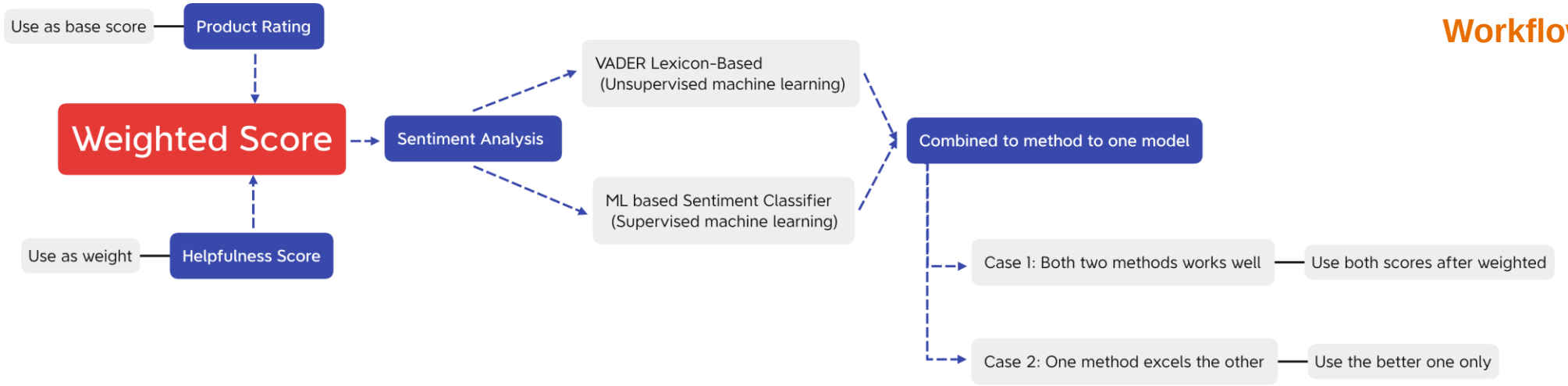
ML based Sentiment Classifier (Supervised)



Try different Model and Tokenizing Method combinations

PRO: Accuracy is likely to be higher

Workflow



Sentiment Analysis: Unsupervised and Supervised Methods

VADER Lexicon-Based (Unsupervised)

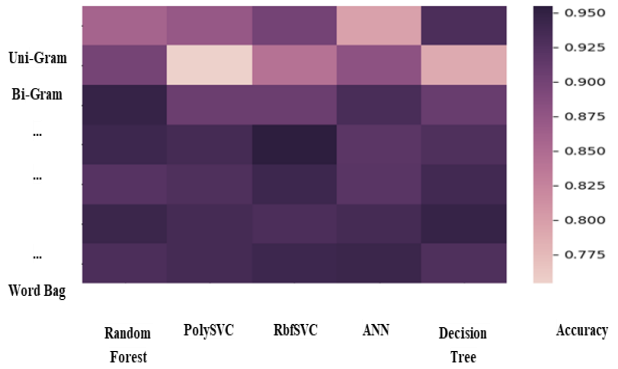
Use Valence Aware Dictionary and sEntiment Reasoner (VADER) lexicon to tell the polarity (positive /negative) and intensity

Polarity

Intensity

PRO: No human-created labels needed

ML based Sentiment Classifier (Supervised)



Try different Model and Tokenizing Method combinations

PRO: Accuracy is likely to be higher

Workflow