Financial Audio Analysis:
Leveraging Speech for Sentiment Classification and Predictability

Springboard DSC - Capstone Project 3

John Arancio

Can financial sentiment analysis be improved using speech audio?

- Sentiment good indicator of market momentum
  - Price action
- NLP sentiment improvement with audio
  - More accurate?
  - NLP sentiment for annotation/labeling
- Audio provides a different way to interpret sentiment
  - Can increase confidence in prospective decisions
  - Signal confluence

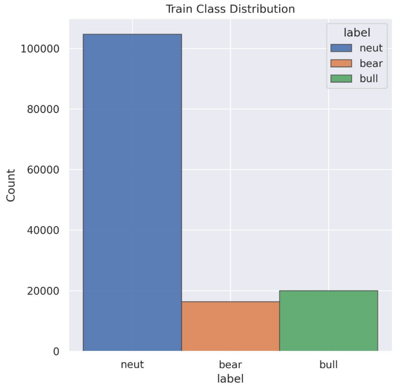
Approach

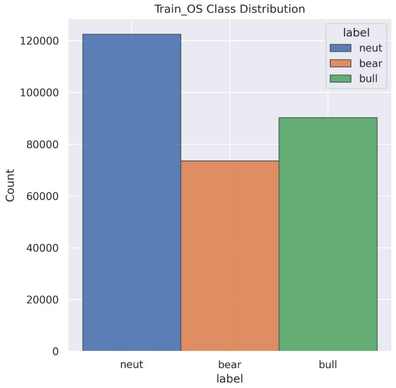
Ensemblement + Wrangling

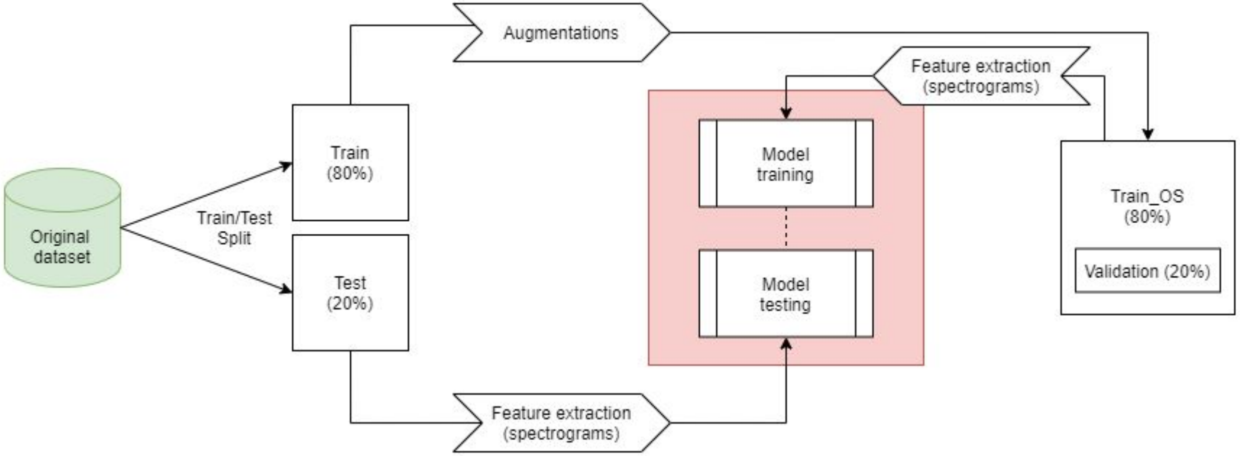
- Web scraped YouTube videos
  - Extract audios
- Transcribe audios
- Segmentation
  - By sentence uttered
- Labeling segments
  - NLP transformer
  - FinBERT
- Padding and reformatting
  - 16-bit, mono, 16000 Hz
  - 8 seconds length

Exploratory Data Analysis

- Class distribution
- Audio properties count (duration, sample rates etc.)
- View waveplots, spectrograms







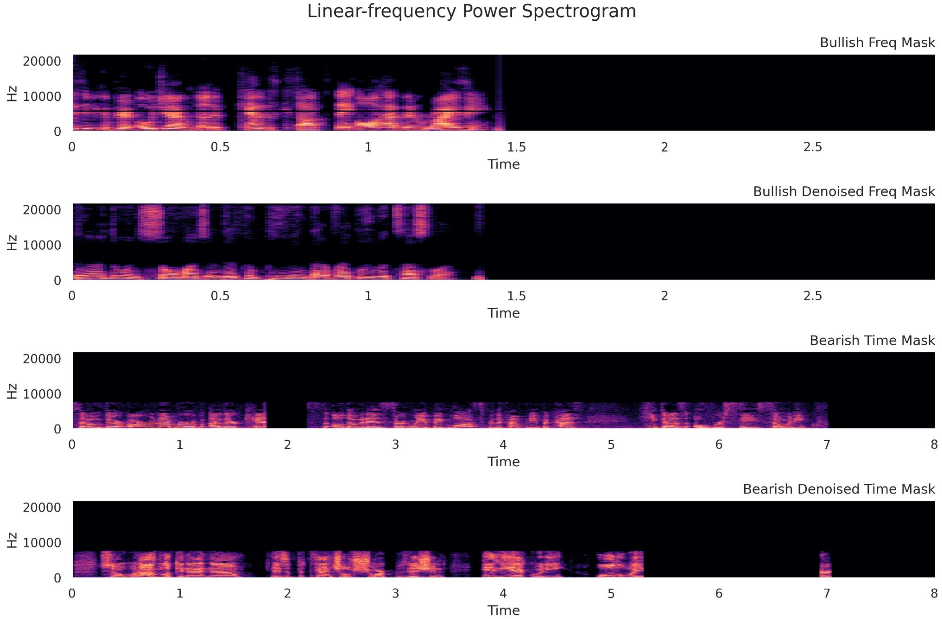
- Split original dataset
- Augmented train set
    - Oversample

- Oversampled dataset training/validation
- Evaluate original test set

Approach - Train/test split

Augmentations

- Sample denoising
- Sample shifting
- Sample frequency mask
- Sample time mask

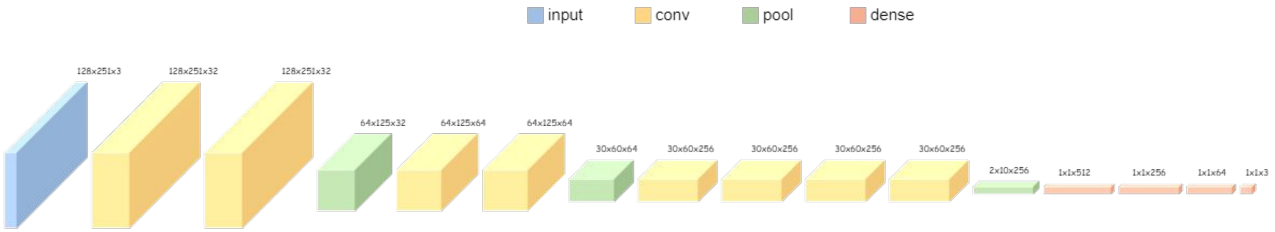


Baseline Modeling

- Logistic Regression
- Support Vector Classifier (SVC)
- XGBoost

LOG. REGRESSION	0	1	2	W.AVG	SVC (RAPIDS)	0	1	2	W.AVG	SVC (SKLEARN)	0	1	2	W.AVG	XGBOOST	0	1	2	W.AVG
PRECISION	0.4	0.38	0.48	0.43	PRECISION	0.94	0.9	0.92	0.92	PRECISION	0.92	0.9	0.92	0.91	PRECISION	0.7	0.56	0.44	0.55
RECALL	0.27	0.45	0.51	0.43	RECALL	0.95	0.93	0.89	0.92	RECALL	0.95	0.93	0.88	0.91	RECALL	0.16	0.2	0.89	0.47
F1-SCORE	0.32	0.41	0.5	0.42	F1-SCORE	0.94	0.91	0.9	0.92	F1-SCORE	0.93	0.92	0.9	0.91	F1-SCORE	0.26	0.29	0.58	0.4
ACCURACY	0.43				ACCURACY	0.92				ACCURACY	0.91				ACCURACY	0.47			

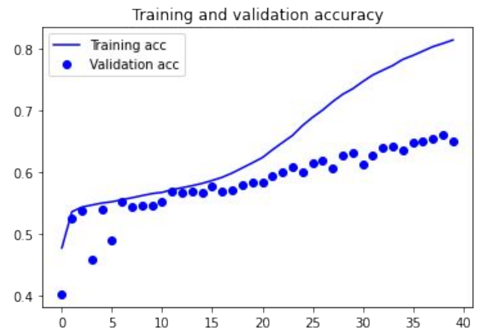
Extended Modeling

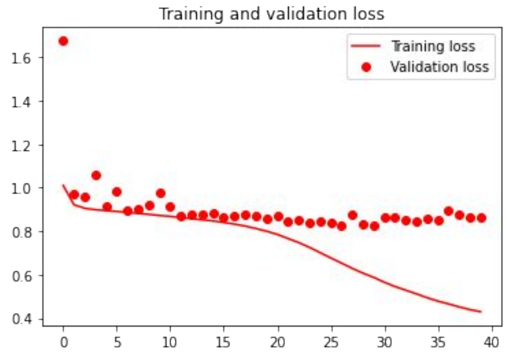


CNN VERSION	TEST LOSS	TEST ACCURACY
v1 (Shallow + tuned)	0.9341	0.5358
v2 (Larger filters + depth)	0.8658	0.5871
v3 (Depth + dropout)	0.8323	0.6192
v4 (Depth + dropout - FC dropout)	0.8134	0.6345

- 4 CNN models
  - Depth added with each version
- V4
  - Max depth
  - Dropout regularization
  - Removed FC dropout

- Inconsistent results across models
    - SVC performed the best
        - F1(weighted avg): 0.92
        - Precision(weighted avg): 0.92
        - Recall(weighted avg): 0.92
    - Logistic Regression performed the worse
        - F1(weighted avg): 0.42
        - Precision(weighted avg): 0.43
        - Recall(weighted avg): 0.43
- CNN v4 has potential
    - Test loss: 0.8134
    - Test accuracy: 0.6345
        - Overfitting (epochs 10-15)
        - Learning rate, no change in training





Findings

Why are our results inconsistent?

- Possibilities:
  - To many augmentations
  - Unbalanced dataset
  - Mel-spectrograms not intricate enough for models to properly generalize

Conclusions

- Mel-spectrograms useful for SVC
- There are useful properties within each class
- Automated annotation process successful
  - Each class was represented by audio properly
- Sentiment can be classified using audio
- Re-evaluate augmentations
- Re-evaluate dataset balancing

Recommendations

1. Use classification results metrics to improve model accuracy
2. Can use SVC to classify audio sentiment as bullish, bearish and neutral
    a. Use model as confluence for decision making
    b. Can fine-tune with more data
3. Use the findings from our EDA to better understand the data collected
4. Try different algorithms comparable to SVM, like “DecisionTree”