

ABSCHLUSSPRÄSENTATION
06.07.2021

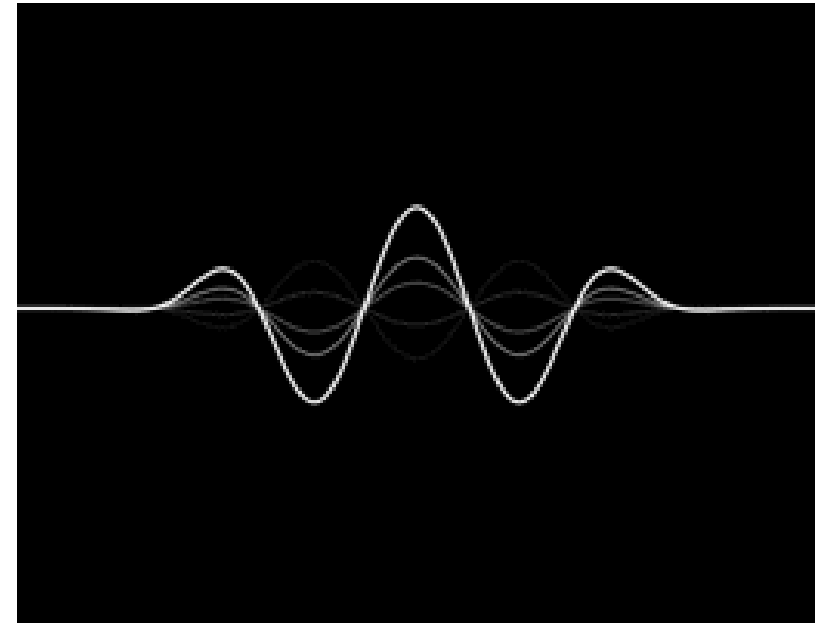
Speech Emotion Recognition

AirTelligence
Philipp Becht, Yannik Hubrich und Simon Wrigg

Inhaltsverzeichnis

- Einführung in die Thematik & Motivation
- Herangehensweise
- Verwendete Technologien & Bibliotheken
- Präsentation der Ergebnisse
- Trainingsprozess
- Kritische Bewertung der Ergebnisse

Einführung in die Thematik & Motivation



Einführung in die Thematik & Motivation

- Verbale Kommunikation zwischen Menschen mithilfe der Sprache
 - Linguistik: Gesprochenes Wort
 - Paralinguistik: Merkmale, die Aufschluss über die Verfassung des Sprechers geben
 - Experiment von Alfred Mehrabian: **7-38-55-Regel**
-
- **Ziel:** Vorhersage und Erkennung der zugrundeliegenden Emotionen einer Audiodatei
-> Es soll keine Sprache, sondern Emotion in der Sprache erkannt werden



Herangehensweise



Herangehensweise

- IDE: Google Colab
- Auswahl und Analyse der Daten
- Einschränkung der verschiedenen Klassen (Emotionen)
- Features von Audiodateien erkennen und auswerten

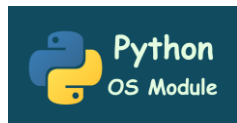


Verwendete
Technologien und
Bibliotheken

Verwendete Technologien und Bibliotheken



Wave





Erster Eindruck der Daten

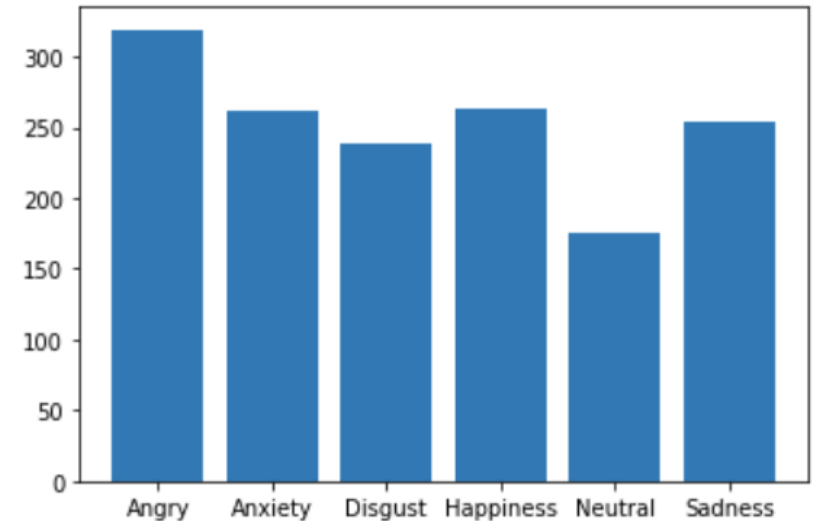
Erster Eindruck der Daten

- Verteilung der Daten über die verschiedenen Klassen
- Insgesamt 1510 Instanzen
- Stichproben: Qualität der Audiodateien

- Beispiel – Struktur der Daten:

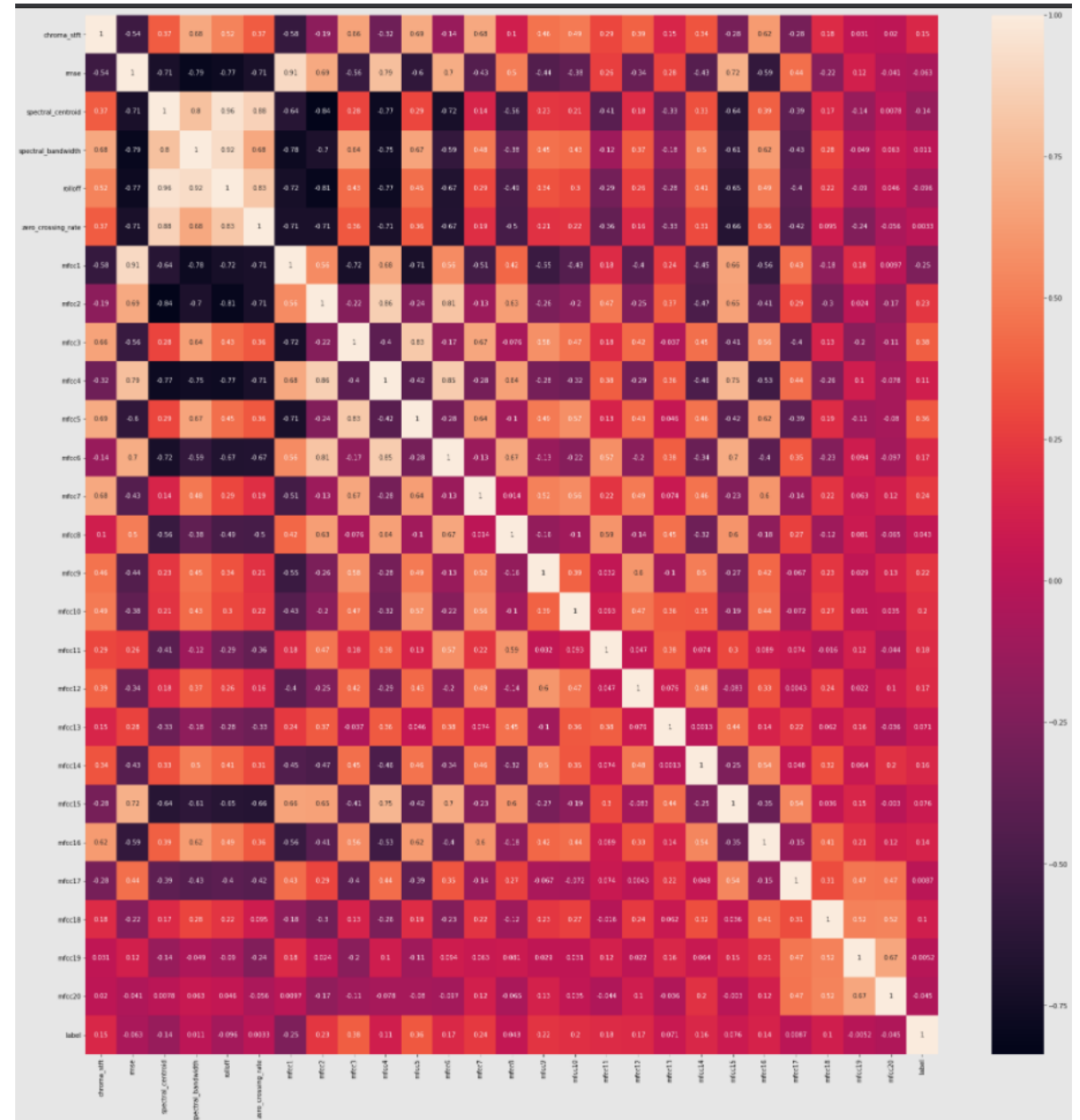
02-01-06-01-02-01-12.mp4

Vocal Channel, Emotion, Emotional Intensity, Statement, # Repetition, Actor (female/male)



Correlation Matrix

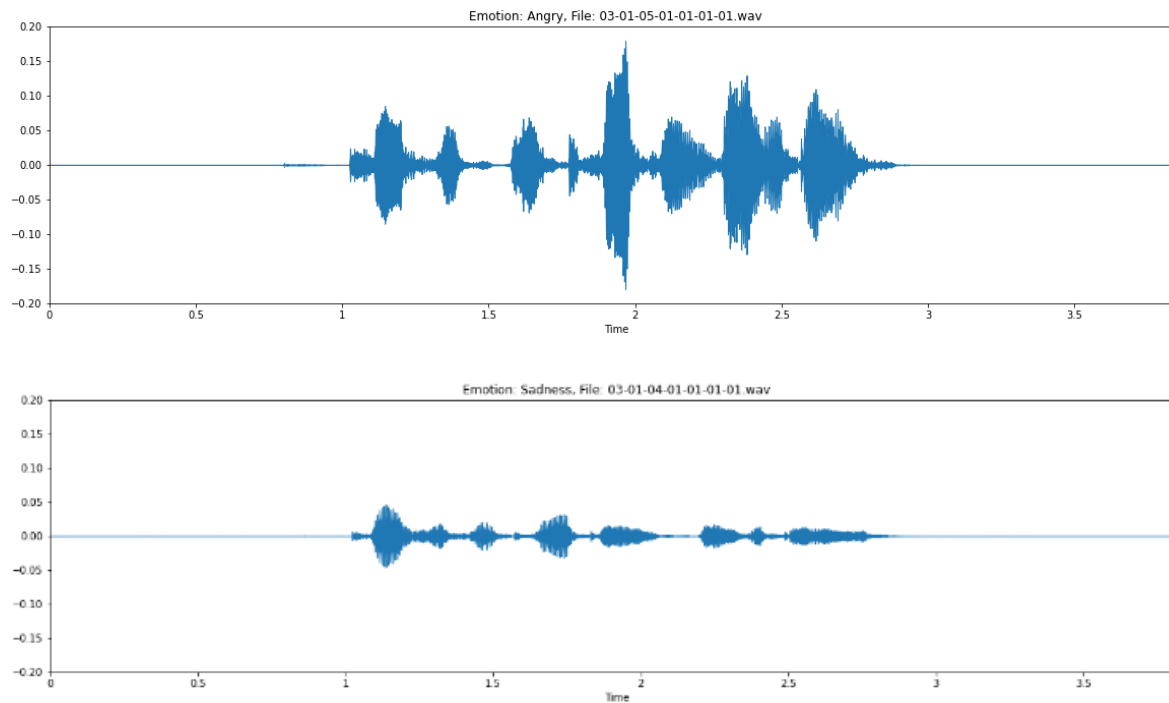
- Untersuchung, ob gewisse Features in besonderem Maße zum Label in Beziehung stehen





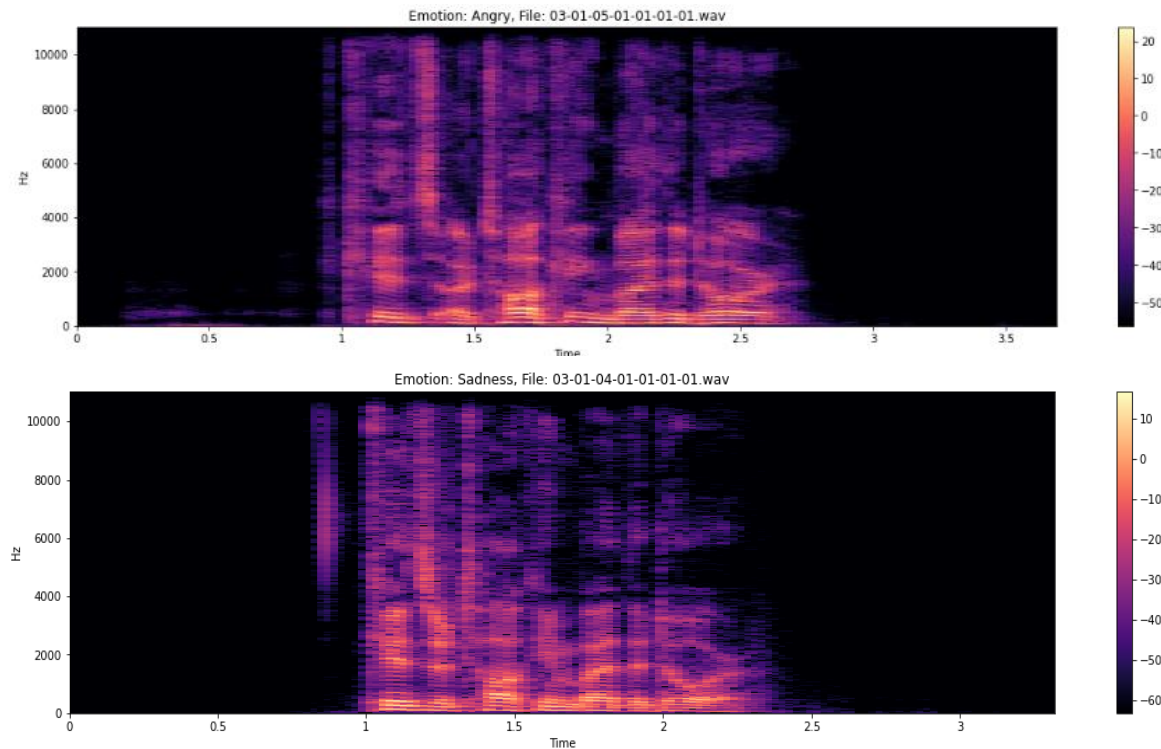
Präsentation der Ergebnisse

Soundwave Plots



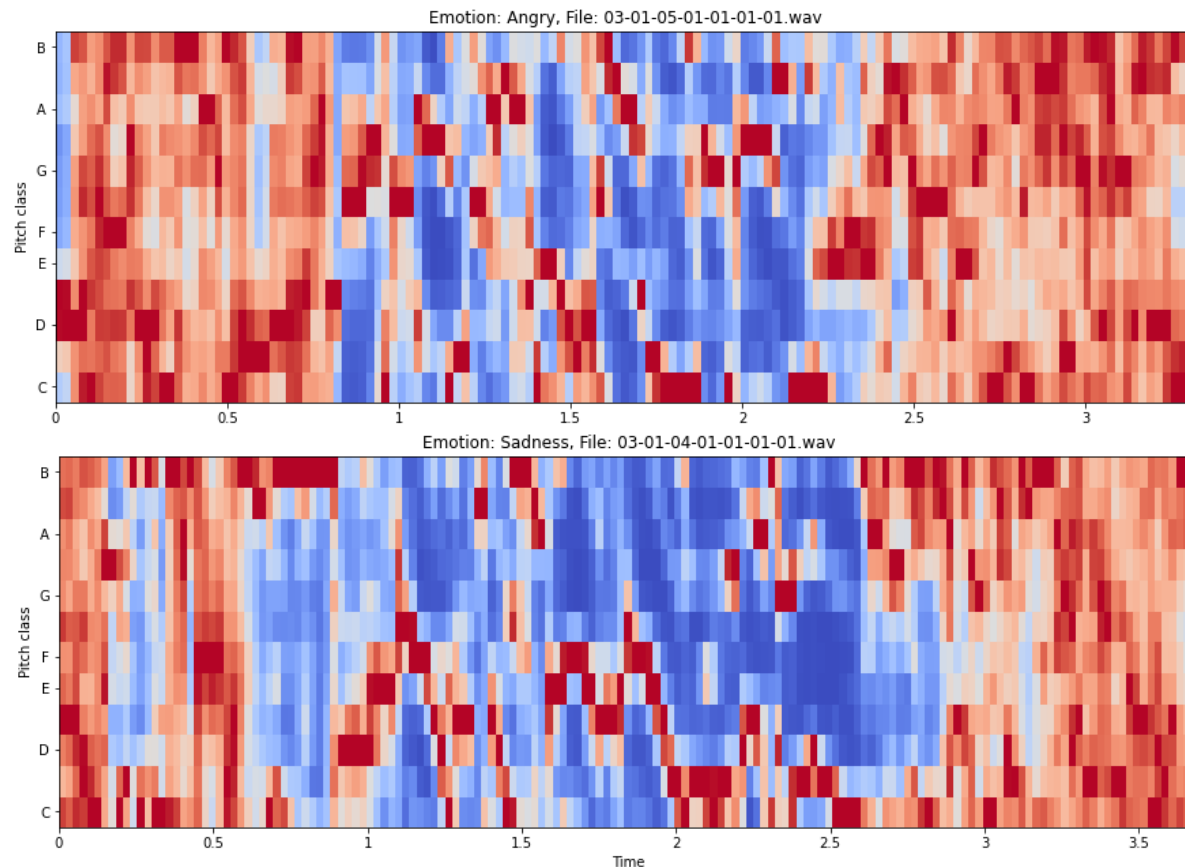
- Visualisierung der **Lautstärke einer Audiodatei** über einen bestimmten Zeitraum
- Klare **Unterschiede der Soundwaves** bei verschiedenen Emotionen erkennbar
- Angry: starke, laute Stimme
- Sadness: ruhige Stimmlage

Spektrogramme



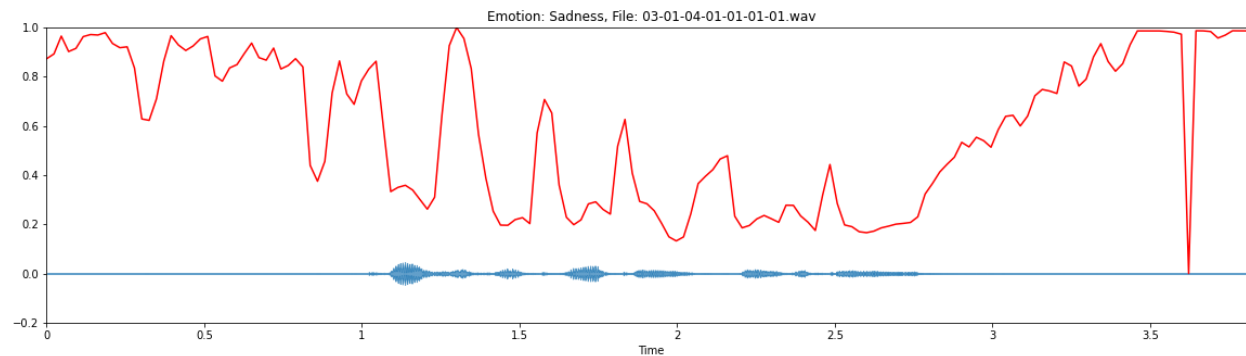
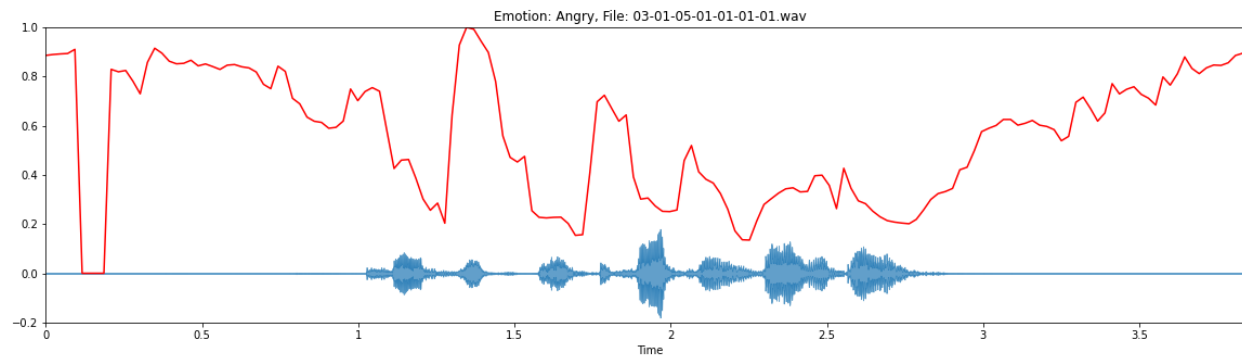
- Visualisierung des **Spektrums der Frequenzen** über einen bestimmten Zeitraum
- Auskunft über Emotionslage der SprecherInnen
- Höherer Frequenzbereich bei der Emotion Angry ist mehr ausgefüllt, als bei Emotion Sadness

Chromagramm



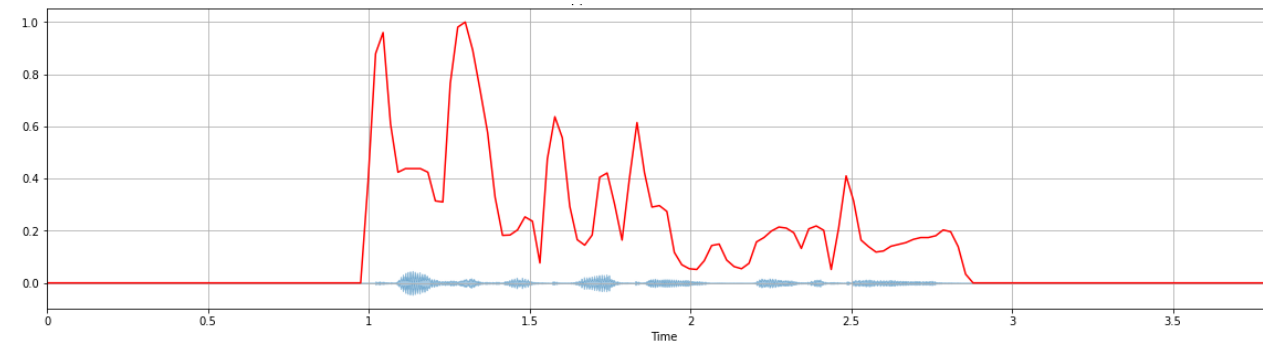
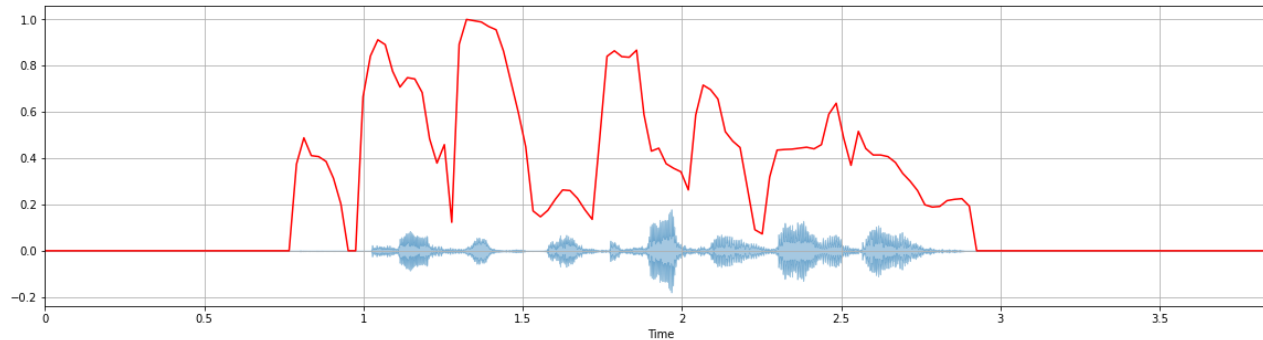
- Chromagramme analysieren die Tonhöhe von gesprochenen Ausschnitten
- Bei der Emotion Angry werden höhere Töne erreicht, als bei dem Ausschnitt in der Emotion Sadness

Spectral Centroids



- Das spektrale Zentroid gibt an, bei welcher Frequenz die Energie eines Spektrums zentriert ist

Spectral Rolloff



- Der Spectral Rolloff ist die Frequenz, unterhalb der ein bestimmter Prozentsatz der gesamten spektralen Energie liegt
- Hier: 85 %


MFCC

(Mel-Frequency Cepstral Coefficients)

- Werden zur automatischen Spracherkennung verwendet
- Modelliert die Eigenschaften der menschlichen Stimme
- Es beschreibt das Frequenzspektrum zusammen mit der wahrgenommenen Tonhöhe und eignet sich somit ebenfalls für die Untersuchungen der Daten.

Erkenntnisse und Bewertung der Visualisierungen

- Bestätigung, dass der Datensatz von ausreichender Qualität ist
- Visualisierungen zeigen Unterschiede in den Audiodateien
- Notwendigkeit der Aneignung eines tiefgreifenden Verständnisses von unserer Stimme und des gesprochenen Wortes



Erstellung einer .csv-Datei für das Training

- csv-Datei mit 26 Features und ein Label
- Features sind teilweise Mittelwerte oder andernfalls tatsächliche Werte
- Zeitpunkt- oder zeitraumbezogen
- Aufsplittung der csv-Datei in einen Trainings- und Testdatensatz mithilfe von scikit-learn



Anwendung von Lernalgorithmen

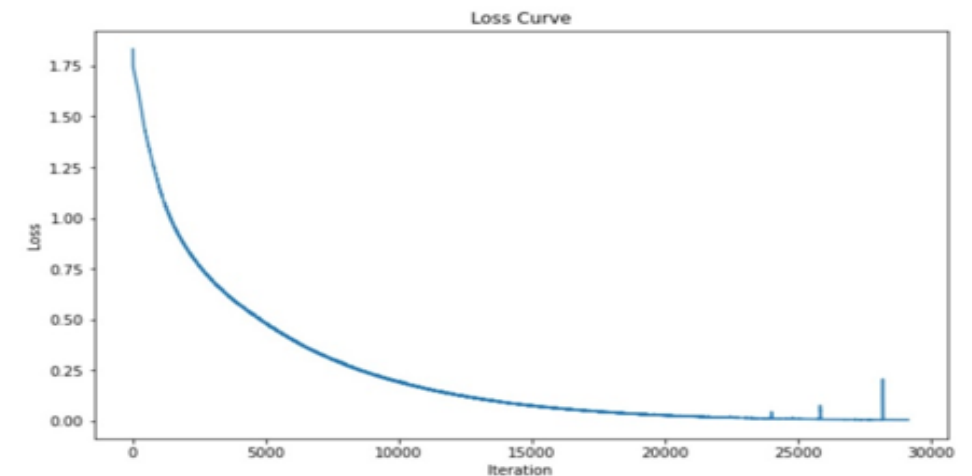
- Entscheidung fiel auf einen Multi Layer Perceptron Classifier (MLP) über die Library Keras
- Voraussetzung: Konvertierung der Audiodateien und Extraktion von Features zu einer .csv-Datei
- Unterschiedliche Initialisierungen des MLP Classifiers (circa 10 Durchläufe in unterschiedlichen Konfigurationen)
- Weitere Modelle wie ähnliche neuronale Netze oder ein KNN-Classifier führten zu schlechteren Ergebnissen

Ergebnisse des Trainings I

- MLP-Classifier
- Verschiedene Einstellungen der Hyperparameter führten zu unterschiedlich guten Ergebnissen (Accuracy auf den Testdaten)
- 29000 Iterationen, 300 Hidden Layer
- Mehrere Durchläufe führten zu einer maximalen Accuracy von **65%**

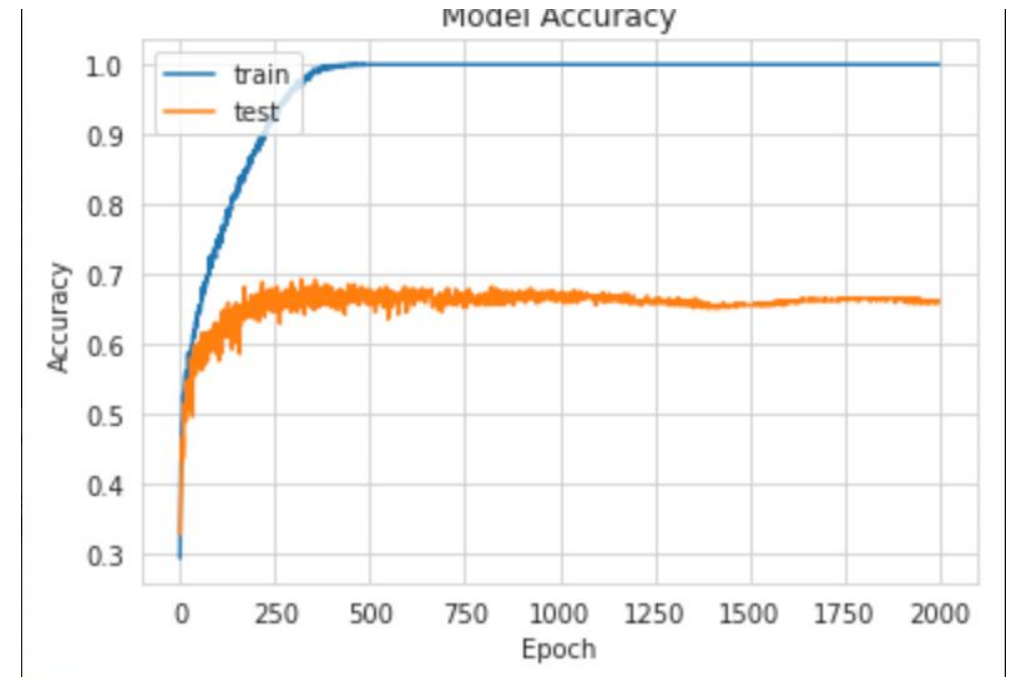
Training loss did not improve more than tol=0.000000 for 1000 consecutive epochs. Stopping.
Training finished after: 10.2 minutes
Number of iterations: 29137
Number of iterations no change: 1000
Model accuracy: 65.23%

	precision	recall	f1-score	support
0	0.81	0.71	0.75	65
1	0.61	0.74	0.67	46
2	0.65	0.55	0.60	51
3	0.49	0.55	0.52	47
4	0.76	0.68	0.72	47
5	0.61	0.67	0.64	46
accuracy			0.65	302
macro avg	0.65	0.65	0.65	302
weighted avg	0.66	0.65	0.65	302



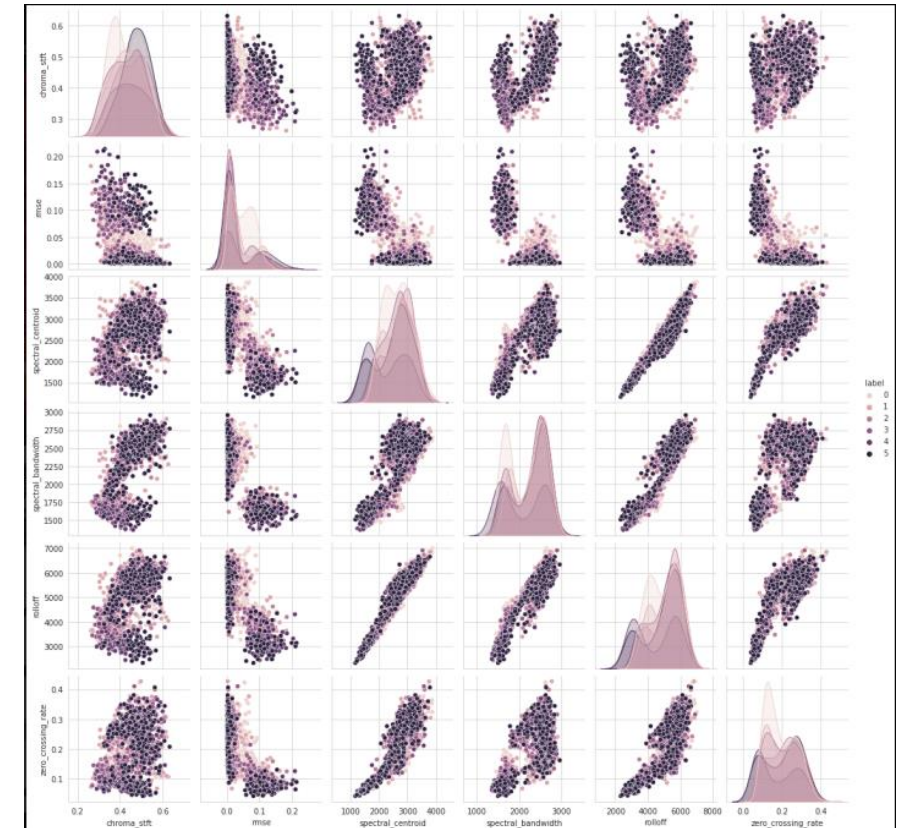
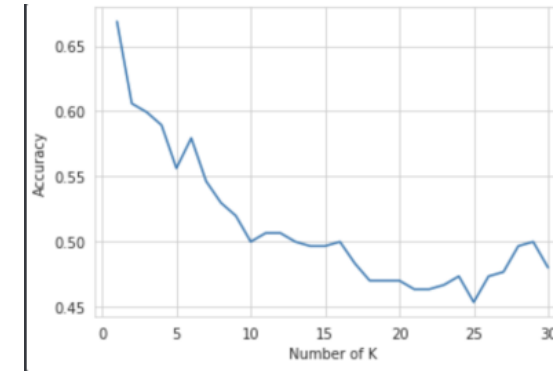
Ergebnisse des Trainings II

- MLP-Classifizier mithilfe von Keras (→ unterschiedliche Library, welche deutlich effizienter ist)
- 2000 Iterationen, 150 Neuronen, Activation Function: tanh
- Führt zu einer letztendlich zufriedenstellenden Accuracy von **69%**



Ergebnisse des Trainings III

- Training KNN-Classifier
- Initialisierung: wäre bei $k=1$ am besten → „Random“
- Eignet sich nicht wirklich für die Klassifikation, siehe auch Grafik





Erkenntnisse aus dem Training

- Die Konfiguration des Lernalgorithmus hat einen erheblichen Einfluss auf die Performance des Modells
- Herausforderung, die optimale Konfiguration (bspw. Activation Function, Anzahl der Hidden Layer, Toleranz) zu finden
- Ein niedriger Loss bedeutet nicht gleichzeitig, dass die Accuracy optimal ist



Kritische Bewertung der Ergebnisse

Kritische Bewertung der Ergebnisse

- Mehrere Versuche, die Lernalgorithmen optimal zu initialisieren
- Tendenz zum Underfitting trotz Anpassung der Hyperparameter
- Die Einstellung der Hyperparameter hat entscheidenden Einfluss auf die Güte des Modells
- Zeit- und ressourcenaufwändiges Training

Kritische Betrachtung des Datensatzes

- Zwei unterschiedliche Datensätze

PRO	CONTRA
<ul style="list-style-type: none">- Verschiedene Sprachen (DE/EN)- Annahme, dass dies zu einer Verbesserung der Robustheit des Modells auf neuen Daten führt	<ul style="list-style-type: none">- Verschiedene Sprachen (DE/EN)- Unterschiedlicher Aufbau der Sätze in den Audiodateien- Annahme, dass dies zu einer Verschlechterung der Accuracy führt, bzw. zu Underfitting auf den Trainingsdaten



Vielen Dank für Eure
Aufmerksamkeit

Datensätze

- <https://www.kaggle.com/nilanshk/emotion-classification-speech>
- <https://www.kaggle.com/piyushagni5/berlin-database-of-emotional-speech-emodb>