Technische Universität Berlin

Fakultät I – Geisteswissenschaften
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation

Exposé

Masterarbeit

Musikemotionserkennung durch Deep Learning auf Grundlage von audio- und textbasierten Informationen

vorgelegt von:

Philipp Scholze 338844

1. Betreuer: Roman Gebhardt
2. Betreuer: Prof. Dr. Stefan Weinzierl

Technische Universität Berlin, Fakultät I – Institut für Sprache und Kommunikation,
Fachgebiet Audiokommunikation
Berlin, 8. März 2021

Inhaltsverzeichnis

Abstract

Durch das rasante Wachstum von digitalen Musikbibliotheken kann sich die Organisation und Bereitstellung dieser Daten problematisch gestalten. Es bedarf neuer Methoden, diese enormen Datenmengen zu handhaben. Abhilfe kann hier unter anderem durch Musikempfehlungsalgorithmen geleistet werden. Dabei stellt die automatische Erkennung der von Musik transportierten Emotionen (Music Emotion Recognition, MER) einen interessanten Ansatz dar. Aufgrund des vielschichtigen Charakters von Musik wurden in den vergangen Jahren vermehrt kombinierte Ansätze untersucht, welche sich sowohl auf das Audiosignal als auch auf die Lyrics der Musikstücke in Textform gestützt haben. Neben klassischen, auf Feature Engineering basierenden Machine-Learning-Ansätzen konnten in jüngster Zeit innerhalb der MER vor allem durch den Einsatz von künstlichen neuronalen Netzwerken deutlich verbesserte Ergebnisse erzielt werden. Die unterschiedlichen Ansätze in der Literatur erfordern jedoch eine Gegenüberstellung und Einordnung, welche im ersten Teil dieser Masterarbeit in Form einer systematischen Übersichtsarbeit geleistet werden soll. Im zweiten Teil der Arbeit soll auf Grundlage der Resultate des ersten Teiles die Entwicklung eines eigenen Emotionsklassifikators im Mittelpunkt stehen. Die Netzwerkarchitektur soll dabei auf Deep-Learning-Methoden wie Convolutional Neural Networks (CNNs) aufbauen und mit bereits bestehenden Architekturen verglichen werden.

1 Einleitung und Fragestellung

Musik ist ein integraler Bestandteil unserer Gesellschaft und allgegenwärtig. So begleitet uns Musik in den meisten alltäglichen Situationen: wir hören Musik morgens am Frühstückstisch, beim Autofahren, im Supermarkt während wir einkaufen, zum konzentrierten Lernen und beim Sport machen (Y.-H. Yang & Chen, 2012). Musik ist aber vor allem auch durch den Aufstieg von Streaming-Anbietern wie Spotify, Apple Music oder Deezer auf mobilen Endgeräten heutzutage nahezu unbegrenzt und jederzeit verfügbar (Schedl et al., 2014, S. 129). Das enorme und stetige Wachstum dieser digitalen Musikbibliotheken wurde in den letzten Jahren nicht nur durch kompakte Audioformate wie MP3, sondern auch durch das Internet maßgeblich begünstigt und beschleunigt. Dabei stellt uns die Organisation und Bereitstellung solch großer Datenmengen an Musik vor unbekannte Probleme und Herausforderungen (Wieczorkowska et al., 2006). Ansätze, welche sich mit der Lösung dieser Probleme beschäftigen, werden im Allgemeinen als Music Information Retrieval (MIR) bezeichnet und sind aktuell Gegenstand intensiver Forschungsbemühungen, sowohl im akademischen als auch im industriellen Rahmen (Casey et al., 2008, S. 669). So werden im MIR insbesondere aussagekräftige Merkmale aus Musikstücken extrahiert, welche unter anderem für kontextbasierte Musiksuchen, Musikempfehlungsalgorithmen oder das Durchsuchen von großen Musikbibliotheken genutzt werden können (Schedl et al., 2014, S. 128).

Ein außerordentlich interessantes, aber mindestens genauso komplexes Merkmal von Musik sind die von ihr transportierten Emotionen oder Stimmungen[1]. Beispielsweise wird Musik von Pratt (1952) als “Sprache der Emotionen” bezeichnet. Mit der

[1]Die Begriffe Emotion und Stimmung werden in der einschlägigen Literatur und auch in dieser Arbeit synonym verwendet.

Music Emotion Recognition (MER) hat sich dementsprechend ein Teilgebiet des MIR gebildet, welches sich überwiegend mit der maschinellen Untersuchung des emotionalen Gehalts von Musikstücken auseinandersetzt. Die Komplexität der MER wird durch ihre Interdisziplinarität deutlich: So sind einerseits Kenntnisse der Signalverarbeitung und des Machine Learnings (ML) notwendig. Andererseits spielen für die MER neben der auditiven Wahrnehmung auch Psychologie und Musiktheorie eine ebenso bedeutsame Rolle (Kim et al., 2010). Die Multimodalität von Musik ist außerdem ein weiterer Faktor, der für die automatisierte Emotionserkennung berücksichtigt werden muss. Neben dem reinen Audiosignal können auch Partituren, Texte (Lyrics), Bilder (Albumcover) oder Gestik (Performer/Interpret:in) mögliche Ausdrucksformen von Musik sein (Schedl et al., 2014, S. 130). Entscheidende Fortschritte in der MER konnten in den letzten Jahren vor allem durch den Einsatz von Machine-Learning-Strategien erzielt werden: hierbei wurden sowohl klassische Ansätze wie das Feature Engineering (FE) als auch moderne Ansätze mit künstlichen neuronalen Netzwerken (artificial neural networks, ANNs) verfolgt. Verfahren, welche die Emotionserkennung entweder nur auf die Analyse des Audiosignals (Wieczorkowska et al., 2006; Malik et al., 2017; Du, Li & Gao, 2020) oder der Lyrics in Textform (Malheiro, Oliveira et al., 2016; Parisi et al., 2019) basieren, konnten bereits gute Ergebnisse aufzeigen. Aufgrund des multimodalen Charakters von Musik scheinen jedoch Methoden, welche die Analyse von Audiosignal und Lyrics kombinieren, eine bessere Abbildung der Realität zu liefern als rein audio- bzw. textbasierte Verfahren. So gab es in der Vergangenheit einige Bemühungen, die automatisierte Emotionserkennung mithilfe kombinierter Ansätze und Feature Engineering zu untersuchen (Malheiro et al., 2013; Laurier et al., 2008; X. Hu & Downie, 2010). Hier konnten ebenfalls gute Resultate erzielt werden. Weiterhin konnten jüngere Publikationen zeigen, dass unter der Verwendung von ANNs bezüglich der MER noch bessere Ergebnisse hervorgebracht werden können (Delbouys et al., 2018; Jeon et al., 2017).

Allerdings stellt die kombinierte, automatisierte Emotionserkennung mit ANNs einen Bereich dar, der bisher noch nicht besonders umfassend erforscht wurde. Ebenso fehlt es an einer ausführlichen Gegenüberstellung von audio- und textbasierten Ansätzen sowie ihrer Kombination, die entweder auf Feature Engineering oder auf künstlichen neuronalen Netzwerken aufbauen. An dieser Stelle soll diese Masterarbeit ansetzen. Entsprechend möchte ich im ersten Teil der Arbeit die bereits bestehenden Forschungsergebnisse miteinander vergleichen und einordnen. Außerdem soll evaluiert werden, unter welchen Rahmenbedingungen diese bereits untersuchten Ansätze nützlich für zukünftige Anwendungen sein können. Anschließend soll im zweiten Teil der Arbeit ein eigener Emotionsklassifikator entwickelt werden, welcher die transportierte Stimmung eines Musikstückes anhand der Kombination von Audiosignal und Lyrics in Textform analysieren und vorhersagen kann. Dieser Klassifikator soll auf der Grundlage von realen Daten unter der Verwendung von künstlichen neuronalen Netzwerken aufgebaut werden. Abschließend soll der entwickelte Emotionsklassifikator mit dem aktuellen Forschungsstand verglichen und diskutiert werden, sowie ein Ausblick für weitere Forschungsansätze gegeben werden.

2 Stand der Forschung

Die systematische Erforschung von Emotionen in der Musik wird seit Anfang des 20. Jahrhunderts vermehrt verfolgt. So untersuchte Kate Hevner bereits 1936 den affektiven Charakter von verschiedenen Musikstücken (Hevner, 1936). Aufgrund der Mehrdeutigkeit des Begriffs „Emotion“ existieren jedoch verschiedene Emotionsmodelle, die sich an einer einheitlichen Darstellung von Emotionen versuchen. Die in der Literatur verwendeten Emotionsmodelle lassen sich dabei in kategorische und dimensionale Modelle unterteilen (Malik et al., 2017). In kategorischen Modellen werden Emotionen über einzelne Worte oder durch Gruppen von Worten (sogenannten Clustern) abgebildet. Eine vergleichsweise einfache Beschreibung bieten die Basisemotionen, die von Ekman (1992) vorgeschlagen wurden. Dabei wird von einer beschränkten Anzahl von diskreten Emotionen ausgegangen, mit denen der Mensch evolutionsbedingt ausgestattet sei. Bezüglich ihrer psychologischen, physiologischen und verhaltensbezogenen Ausprägungen werden diese Basisemotionen als unabhängig voneinander betrachtet, wobei jede Emotion durch eine spezifische Aktivierungen im zentralen Nervensystem ausgelöst werde (Posner, Russell & Peterson, 2005). So identifizieren Ekman und Friesen (2003, S. 22) Fröhlichkeit (eng. happiness), Traurigkeit (sadness), Angst (fear), Ekel (disgust), Wut (anger) und Überraschung (surprise) als die sechs Basisemotionen, die durch Gesichtsmimik ausgedrückt werden können, wobei weitere Emotionen wie Scham (shame) und Begeisterung (excitement) als Mischung der sechs Basisemotionen verstanden werden können.

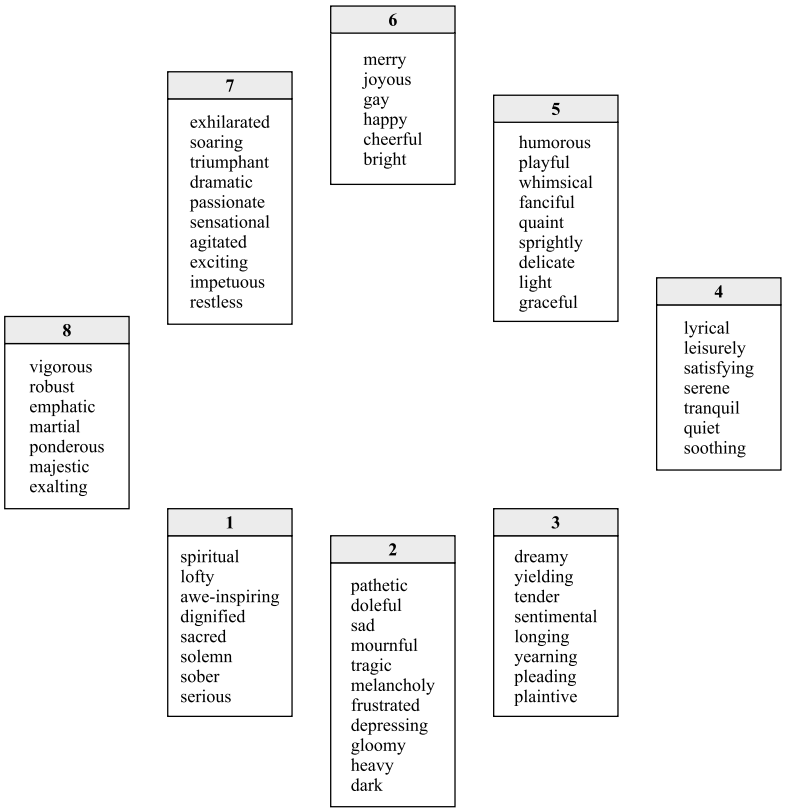

Abbildung 1: Adjektivkreis nach Hevner (1936).

Aufgrund der recht allgemeinen Beschaffenheit der Basisemotionen wurden weitere, bereichsspezifische Modelle zur Beschreibung von Emotionen in der Musik erarbeitet. Der von Hevner (1936) entwickelte Adjektivkreis (adjective circle) in Abbildung 1 stellt dabei ein bekanntes Beispiel für ein bereichsspezifisches, kategorisches Emotionsmodell dar. So wurden 66 englische Adjektive in acht Gruppen angeordnet, wobei sich die Anzahl der Adjektive pro Gruppe unterscheiden können. Das von X. Hu und Downie (2007) beschriebene Emotionsmodell zeigt dagegen einen neueren, datenbasierten Ansatz. Hier wurden Emotionscluster (siehe Abbildung 2) anhand der Auswertung von online verfügbaren Metadaten[2] erstellt. Dieses Modell wurde inzwischen auch für den MIREX Mood Classification Task[3] adaptiert (Panda, 2019, S. 28).

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Rowdy	Amiable /	Literate	Witty	Volatile
Rousing	Good natured	Wistful	Humorous	Fiery
Confident	Sweet	Bittersweet	Whimsical	Visceral
Boisterous	Fun	Autumnal	Wry	Aggressive
Passionate	Rollicking	Brooding	Campy	Tense/anxious
	Cheerful	Poignant	Quirky	Intense
			Silly	

Abbildung 2: Cluster-Modell nach X. Hu und Downie (2007).

Dimensionale Modelle hingegen verfolgen allgemein den Ansatz, Emotionen in einem kontinuierlichen, mehrdimensionalen Raum darzustellen. Häufig wird hier das sogenannte „Circumplexmodell“ von Russell (1980) verwendet, welches Emotionen in einem zweidimensionalen Raum abbildet, wobei hier Arousal und Valenz die beiden Achsen beschreiben. Unter dem Begriff Arousal wird die Intensität einer Emotion, welche durch einen Stimulus hervorgerufen wird, verstanden. Die Valenz hingegen beschreibt, wie angenehm ein Stimulus empfunden wird (Warriner et al., 2013). Eine grafische Veranschaulichung des Circumplexmodells findet sich in Abbildung 3. Quadrant 1 wird mit „erfreut“ (eng. joyful), der zweite Quadrant mit „zufrieden“ (content) beschrieben. Für Quadrant 3 findet sich die Beschreibung „bedrückend“ (depressing), wohingegen Quadrant 4 die Emotion „würtend“ (angry) darstellt (Parisi et al., 2019). Das Circumplexmodell nach Russell (1980) hat sich inzwischen laut Panda (2019, S. 28) zum dimensionalen Standardmodell in der MER-Forschung entwickelt.

Jedoch betont Panda (2019, S. 37) ebenfalls, dass die Wahl des Emotionsmodells nicht trivial sei und immer vom eigentlichen Forschungszweck abhänge. Bereichsspezifische, kategorische Modelle seien besser für induzierte Emotionen geeignet, während sich wahrgenommene Emotionen günstiger durch Basisemotionen ausdrücken lassen. Dagegen können dimensionale Modelle zwar das Problem der Mehrdeutigkeit von Emotionen beheben. Jedoch können sie auch einen weitaus größeren, rechnerischen Aufwand mit sich bringen und werden eventuell nicht so gut von Probanden in Hörtests verstanden wie einfache kategorische Emotionsmodelle. Neben den verwendeten Emotionsmo-

[2]Die Daten wurden von den Webseiten AllMusicGuide.com, epinions.com und Last.fm bezogen.

[3]MIREX steht für Music Information Retrieval Evaluation eXchange. Hier werden jährlich moderne MIR Algorithmen verglichen im Rahmen der ISMIR (International Society for Music Information Retrieval) Konferenz.

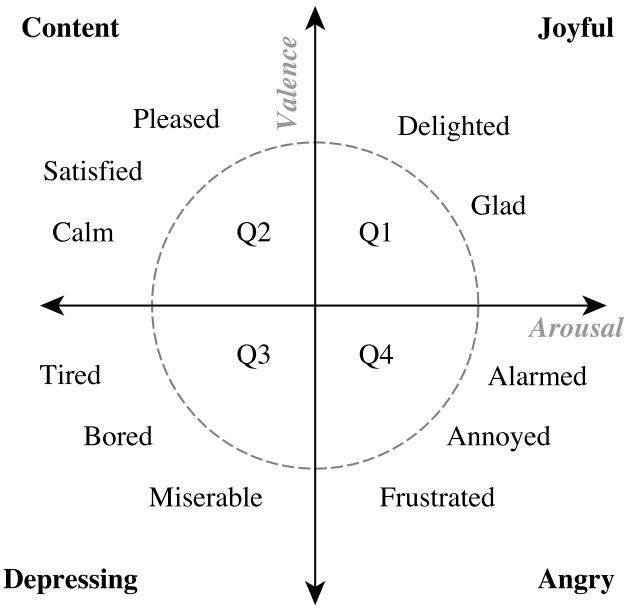

Abbildung 3: Circumplexmodell nach Russell (1980) und Parisi et al. (2019).

dellen muss auch unterschieden werden, ob es sich um eine statische oder dynamische Emotionserkennung handelt. Statische Methoden weisen dabei einem Musikstück als ganzes eine Emotion zu. Bei dynamischen Ansätzen werden die Musikstück jedoch in Segmente aufgeteilt und jedes dieser Segmente wird mit einer Emotion versehen (Du et al., 2020).

2.1 Textbasierte Modelle

Die Verwendung von Lyrics in Textform zur Emotionserkennung stellt eines der beiden, in dieser Arbeit untersuchten, Modelle dar. Es ist anzumerkden, dass die Extrahierung des emotionalen Gehalts von Lyrics durchaus eine Herausforderung sein kann (Kim et al., 2010, S. 259). Mithilfe eines affektiven Wörterbuches konnten Y. Hu, Chen und Yang (2009) unter der Verwendung des Circumplexmodells nach Russell (1980) und eines Datensatzes von 981 chinesischen Musikstücken aus den Genres Pop, Rock und Rap vor allem für positive Valenz- und Arousalwerte ("fröhlich", eng. happy) mit einem F1-Maß von 0,6975 gute Ergebnisse erzielen. Van Zaanen und Kanters (2010) nutzen ebenfalls ein zweidimensionales Emotionsmodell mit Valenz und Arousal als Achsen. So wurden die Lyrics von 5631 Musikstücken analysiert und mit einem k-nearest-Neighbour-Algorithmus (KNN) klassifiziert, wobei eine durchschnittliche Genauigkeit von 77,18% für Arousal und 76,26% für Valenz für k=1 erzielt werden konnte. Hervorzuheben ist, dass Van Zaanen und Kanters (2010) insbesondere die sogenannte Term Frequency Inverse Document Frequency (TF-IDF) für die betrachteten Lyrics berechnet haben, wodurch die relative Bedeutung eines Wortes für eine bestimmte Emotionsgruppe ausgedrückt werden soll. Ein anderer Ansatz wurde von Malheiro, Oliveira et al. (2016) gewählt. Hier sollte berücksichtigt werden, dass sich die Stimmung auch im Verlauf der Lyrics ändern könne, wodurch eine auf Schlüsselwörtern basierende, dynamische Emotionserkennung untersucht wurde. Malheiro, Oliveira et

al. (2016) orientierten sich dabei ebenfalls am Circumplexmodell und konnten anhand von insgesamt 368 Versen aus den Lyrics von 112 Musikstücken ein F1-Maß von 67,7% erreichen. Dieser Werte konnte im selben Jahr von Malheiro, Panda, Gomes und Paiva (2016) mithilfe von Support Vector Machines (SVMs) und der Betrachtung von inhaltlichen Merkmalen wie Bag-of-Words (BOW) in Form von Mono-, Bi- und Trigrammen auf 73,6% verbessert werden. Parisi et al. (2019) setzen in ihrer Publikation auf verschiedene ANN-Modelle, welche auf Deep Learning (DL) basieren. Die Emotionserkennung beschränkte sich dabei auf ein kategorisches Emotionsmodell, welches aus den fünf Basisemotionen Traurigkeit (sadness), Freude (joy), Angst (fear), Wut (anger) und Ekel (disgust) besteht. Das beste F1-Maß (67,1%) wurde von einer Netzwerkarchitektur erzielt, die sich aus dem von Facebook entwickelten Word Embedding *fastText*[4] und einem Long Short-Term Memory Netzwerk (LSTM) mit Aufmerksamkeitsmechanismus zusammensetzt. Obwohl hier ein gutes Ergebnis erzielt werden konnte, muss trotzdem erwähnt werden, dass die fünf Basisemotionen innerhalb des Datensatzes sehr ungleich verteilt sind: so sind lediglich 2,1% der Lyrics als „Ekel“ gekennzeichnet, während 43,9% mit der Emotion „traurig“ versehen wurden. Weiterhin ist die Größe des verwendeten Datensatzes nicht bekannt.

2.2 Audiobasierte Modelle

Ansätze, welche sich auf die Betrachtung des Audiosignals beschränken, konnten ebenfalls gute Ergebnisse bezüglich der MER aufzeigen. In einem auf Feature Engineering basierten Versuch verwendeten Li und Ogihara (2003) auditive Merkmale wie Timbre, Rhythmus und Tonhöhe um Musikstücke in eine von insgesamt 13 Emotionsgruppen einzuteilen. Die verwendeten Emotionsgruppen sind dabei an den Adjektivkreis von Hevner (1936) angelehnt. Untersucht wurden 499 Ausschnitte von Musikstücken mit einer Länge von jeweils 30 Sekunden. Die Musikstücke wurden aus den Genres Ambient, Klassik, Jazz und Fusion bezogen. Mithilfe von SVMs konnte eine Genauigkeit von 45% (F1-Maß) für Emotionserkennung erzielt werden. Wieczorkowska et al. (2006) übernahmen das kategorisches Emotionsmodell von Li und Ogihara (2003), untersuchten aber einen deutlich größeren Datensatz. So wurden ebenfalls 30 Sekunden lange Ausschnitte von 875 Popsongs und klassischen Musikstücken bezüglich 29 Audiodeskriptoren wie der Grundfrequenz, Lautstärke und dem Anteil an geraden und ungeraden Harmonischen ausgewertet. Neben den 13 Emotionsgruppen wurde die Anzahl der Gruppen durch die Einführung von sechs Supergruppen reduziert und ebenfalls analysiert. Für die Emotionsklassifizierung wurde ein KNN-Algorithmus und eine fünffache Kreuzvalidierung (eng. cross validation, CV) angewandt. So konnte für die 13 Emotionsklassen eine Genauigkeit von 27,1% (k=13), für die sechs Supergruppen sogar eine Genauigkeit von 38,6% (k=15) erreicht werden. Diese Ergebnisse würden den Ergebnissen von subjektiven Tests mit Probanden entsprechen. Ähnlich zu Li und Ogihara (2003) untersuchten Lu, Liu und Zhang (2005) akustische Merkmale wie Intensität, Timbre und Rhythmus zur Emotionserkennung. Dabei wurde jedoch ein dimensionales Emotionsmodell verwendet, welches an das zweidimensionale Circumplexmodell von Russell (1980) angelehnt ist. Dabei wurden den vier Quadranten des zweidimensionalen Emotionsmodells folgende Emotionen zugewiesen: Zufriedenheit (contentment), Traurigkeit (depression), Ausgelassenheit (exuberance) und Angst (anxious/frantic). Beim Datensatz handelt es sich um 800 Ausschnitte von 250 klassischen Musikstücken

[4]https://fasttext.cc/, zuletzt abgerufen am 16.12.2020.

mit einer Länge von jeweils 20 Sekunden, wobei die Emotionskennzeichnung manuell von drei Experten erstellt wurden. Als Klassifikator dienten Gaussian Mixture Models (GMM), welche eine durchschnittliche Genauigkeit von 86,3% erreichen konnten. Ein dreidimensionales Emotionsmodell, welches die fünf Basisemotionen „fröhlich“ (happy), „traurig“ (sad), „liebevoll“ (tender), „beänstigend“ (scary/fear) und „wütend“ (angry) anhand der drei Achsen Anspannung (tension, „angespannt“ bis „entspannt“), Aktivität (activity, „wach“ bis „müde“) und Valenz für einen Datensatz von 110 Ausschnitten aus Filmmusik einordnet, wurde von Eerola, Lartillot und Toiviainen (2009) untersucht. Als auditive Merkmale wurden unter anderem die Dynamik, Timbre, Rhythmus und Harmonie verwendet. Eine Regression der partial kleinsten Quadrate (partial least square regression, PLS) liefert dabei die besten Ergebnisse: so konnte ein R^2 von 0,72 für Valenz, 0,85 für Aktivität und 0,79 für Anspannung erzielt werden. Die PLS konnte für die fünf Basisemotionen folgende R^2 Werte erreichen: 0,58 (liebevoll); 0,68 (fröhlich); 0,69 (traurig); 0,70 (wütend) und 0,74 (beängstigend).

Neuere Forschungsansätze haben neben dem klassischen Feature Engineering auch künstliche neuronale Netzwerke untersucht. T. Liu, Han, Ma und Guo (2018) konnten zeigen, dass der Einsatz von Convolutional Neural Networks (CNN) die durchschnittliche Genauigkeit (0.724) der Emotionserkennung im Vergleich zu klassischen Machine Learning Methoden wie den SVM (0.385) deutlich steigern kann. Im von Russell (1980) adaptierten Circumplexmodell teilen T. Liu et al. (2018) die zweidimensionale Valenz-Arousal-Ebene (Abszisse: Valenz, Ordinate: Arousal) wie folgt auf: -45 bis 45 Grad „zufrieden“/„fröhlich“ (pleased/happy), 45 bis 135 Grad „erregt“/„beunruhigt“ (aroused/alarmed), 135 bis 225 Grad „unglücklich“/„traurig“ (miserable/sad), 225 bis 315 Grad „beruhigt“/„müde“ (soothing/tired). Es wurden 744 Ausschnitte von Musikstücke aus dem Free Music Archive[5] mit je 45 Sekunden Länge analysiert, wobei die Ausschnitte noch einmal Segmente von fünf Sekunden Länger unterteilt wurden, um den Datensatz zu vergrößern. Für jedes Segment wurde ein Spektrogramm[6] berechnet und das CNN mithilfe der Spektrogramme trainiert. X. Liu et al. (2017) implementierten ebenfalls CNNs mit Spektrogrammen, allerdings mit einem kategorischen Emotionsmodell bestehend aus 18 verschiedenen Emotionen. Als Grundlage dienten der CAL500[7] Datensatz mit 502 Musikstücken und der CAL500exp[8] Datensatz mit 3223 Musiksegmenten. Für die Emotionserkennung konnte ein durchschnittliches F1-Maß von 0,709 erreicht werden. Auch die Kombination von mehreren Netzwerkarchitekturen zur automatischen Emotionserkennung wurde erforscht. So verbanden Malik et al. (2017) ein CNN mit einem Recurrent Neural Network (RNN), wobei das RNN als bidirektionales LSTM-Netzwerk implementiert wurde. Auch hier wurde Circumplexmodell von Russell (1980) übernommen. Während für den Trainingsdatensatz 431 je 30 Sekunden lange Ausschnitte von Musikstücken verwendet wurden, fungierten 58 Musikstücke in ihrer kompletten Länge als Testdatensatz. Unter der Verwendung von grundlegenden Audiodeskriptoren wie unter anderem den Mel Frequency Cepstral Coefficients (MFFCs), Spectral Centroid und Spectral Rolloff konnte der mittlere quadratische Fehler (Root-Mean-Square Error, RMSE) auf 0,267 für Valenz und auf 0,202 für Arousal ver-

[5]https://freemusicarchive.org/ , zuletzt abgerufen am 07.12.2020.

[6]Durch ein Spektrogramm wird der zeitliche Verlauf des Frequenzspektrums eines Audiosignals mithilfe von Farbkodierung bildlich dargestellt.

[7]https://github.com/yzhaobk/CAL500, zuletzt abgerufen am 07.12.2020.

[8]http://slam.iis.sinica.edu.tw/demo/CAL500exp/, zuletzt abgerufen am 07.12.2020.

ringert werden. Du et al. (2020) konnten die Resultate von Malik et al. (2017) noch einmal deutlich verbessern: so erreichten sie ebenfalls mit einer Kombination von CNN und bidirektionalem LSTM einen RMSE-Wert für die Valenz von 0,06 bzw. 0,07 für Arousal. Allerdings wurde hier zum einen ein deutlich größerer Datensatz verwendet, welcher 1000 westliche Popsongs umfasst[9]. Zum anderen haben Du et al. (2020) sowohl Mel-Spektrogramme als auch Cochleogramme der Musikstücke berechnet. Ein Cochleogramm überführt das Audiosignal in einen multidimensionalen Vektor durch den Einsatz von Gammaton-Filtern und soll so die Informationen, die vom Ohr zum Gehirn gesendet werden, repräsentieren.

2.3 Kombinierte Modelle

Eine der ersten Forschungsbemühungen, welche einen kombinierten Ansatz basierend auf dem Audiosignal und den Lyrics in Textform untersuchte, wurde von D. Yang und Lee (2004) unternommen. In dieser auf Feature Engineering aufbauenden Arbeit wurden für die automatische Emotionsklassifizierung in elf Kategorien sowohl die Lyrics als auch auditive Merkmale wie Tempo oder Spectral Centroid untersucht. Der Datensatz setzte sich dabei aus 145 je 30 Sekunden langen Ausschnitten von Musikstücken zusammen. Es konnte eine Genauigkeit von 82,8% erreicht werden, wobei das kombiniertes Modell nur geringfügig bessere Ergebnisse als ein vergleichbares, audiobasiertes Modell (80,7% Genauigkeit), lieferte. Laurier et al. (2008) kombinierten verschiedene textliche und akustische Merkmale, um die Emotionserkennung für die vier Quadranten des Circumplexmodells vorzunehmen. Neben klangfarblichen, rhythmischen oder tonalen Audiodeskriptoren wurden für die Lyrics vor allem auch Language Model Differences (LMD) untersucht. Mit den LMD konnten für jeden Quadranten die 100 aussagekräftigsten Terme bestimmt werden, welches zu einer besseren Emotionserkennung führte. So konnte die Genauigkeit im Vergleich zur alleinigen Verwendung von Audiomerkmalen vor allem für die Quadranten „fröhlich“ (happy) von 81,5% auf 86,8% und „traurig“ (sad) von 87,7% auf 92,8% gesteigert werden. Für „entspannt“ (relaxed) (von 91,4% auf 91,7%) und „wütend“ (angry) (von 98,1% auf 98,3%) konnte nur eine geringfügige Verbesserung erzielt werden, wobei die Klassifizierungsgenauigkeit hier bereits sehr hoch lag. Ein vergleichsweise großer Datensatz von annähernd 3000 Musikstücken wurde von X. Hu, Downie und Ehmann (2009) analysiert. Hier wurde ein kategorisches Emotionsmodell mit insgesamt 18 Emotionsklassen herangezogen. Neben spektralen Audiomerkmalen wurden die Lyrics durch BOW mit Stemming[10] und TF-IDF-Gewichtung repräsentiert. Wie bereits bei Laurier et al. (2008) wurden die LMD für jede Emotionsklasse bestimmt, wodurch im kombinierten Ansatz von Lyrics- und Audiomerkmalen die höchste Genauigkeit für die Emotionserkennung in 13 von 18 Emotionsklassen erreicht werden konnte. Schuller, Dorfner und Rigoll (2010) verfolgten den Ansatz eines diskreten Valenz-Arousal-Raumes (VA-Raum) für die automatische Emotionserkennung. Dazu wurde ein Datensatz von 2648 Popsongs ausgewertet und der diskrete VA-Raum in zwei unabhängige Klassifizierungsprobleme mit jeweils drei Klassen aufgeteilt. Eine algorithmische Auswahl von textlichen und auditiven Merkmalen führte zu einem System mit einer Genauigkeit von 64,4% für die Valenz- und 60,9% für die Arousalerkennung. Mithilfe von elf Audiodeskriptoren und Unigrammen

[9]Der Datensatz wurde von der Universität Genf bereitgestellt.

[10]Stemming beschreibt die Reduktion eines Wortes auf einen gemeinsamen Wortstamm. So werden zum Beispiel die Wörter „gesehen“ und „sah“ auf den gemeinsamen Wortstamm „sehen“ reduziert.

konnten Malheiro et al. (2013) unter der Verwendung von SVM die Emotionserkennung von fünf Emotionsgruppen von 62,4% (F1-Maß, nur Audiomerkmale) auf 63,9% erhöhen. Dazu wurde ein Datensatz von 764 Ausschnitten von Musikstücken untersucht.

Jüngste Studien haben indes auch Deep-Learning-Methoden (DL) für kombinierte Modelle herangezogen. Einen ersten Ansatz lieferten Jeon et al. (2017) mit ihrem Versuch, eine DL-Architektur für die Emotionserkennung zu verwenden. Die Architektur setzte sich aus einem Convolutional Recurrent Neural Network (CRNN) für die Audiodaten und einem CNN für die Lyrics zusammen, wobei Audiosignale als Mel-Spektrogramme und Lyrics als Wortvektoren repräsentiert wurden. Mit einer Genauigkeit von 80,46% konnte von diesem Modell ein beachtliches Ergebnis erreicht werden. Es ist jedoch anzumerken, dass der hier analysierte Datensatz von 7484 K-Pop Musikstücken nur ein binäres Emotionsmodell (positiv oder negativ) abdeckt. Bhattacharya und Kadambari (2018) entschieden sich für ein kategorisches Emotionsmodell mit fünf Emotionsgruppen. Als Architektur wurde ebenfalls ein CNN für beide Modalitäten verwendet, wobei Mel-Spektrogramme der Audiosignale berechnet und ein 100-dimensionales Word Embedding[11] angewandt wurde. Für zwei verschieden große Datensätze (MIREX Dataset mit 903 Einträgen und Million Song Dataset[12] mit 48476 Einträgen konnte das F1-Maß für die Emotionserkennung auf 66,28% (MIREX) bzw. 69,73% (MSD) gesteigert werden. Delbouys et al. (2018) beziehen ebenfalls einen Teil ihres 18644 Musikstücke umfassenden Datensatzes aus dem MSD. Neben verschiedenen DL-Methoden für Audiosignale und Lyrics werden hier allerdings auch klassische FE-Ansätze untersucht und untereinander verglichen. Auch hier konnte gezeigt werden, dass kombinierte Modelle eine bessere Emotionserkennung liefern als rein text- oder audiobasierte. Obwohl für die Arousaldetektion durch Deep Learning bessere Resultate erzielt werden konnten ($R^2 = 0,235$), waren FE-Ansätze für die Valenzerkennung gleichermaßen performativ wie Deep Learning ($R^2 = 0,219$). Weiterhin konnte herausgefunden werden, dass sich insbesondere die Valenzerkennung verbessert, wenn die Ausgänge des Lyrics- und des Audionetzwerkes früher zusammengeführt werden.

3 Methodisches Vorgehen

Wie bereits in der Einleitung erwähnt, soll diese Masterarbeit zwei Teile umfassen: im ersten Teil soll die bestehende Literatur mit einer systematischen Übersichtsarbeit zusammengefasst werden. Der zweite Teil soll den praktischen Teil abdecken, indem ein eigener Emotionsklassifikator entwickelt werden soll. Entsprechend der beiden Aufgabenteilen der MA ist das methodische Vorgehen ausgewählt.

3.1 Systematische Übersichtsarbeit

Aufgrund der sehr unterschiedlichen Vorgehensweisen in bestehenden Literatur ist es notwendig, diese Arbeiten entsprechend zu kategorisieren, einzuordnen und zu vergleichen. Dies soll im ersten Teil der Masterarbeit in Form einer systematischen Übersichtsarbeit geschehen und soll sowohl für zukünftige Arbeiten, als auch für diese

[11] GloVe-Repräsentation, siehe https://nlp.stanford.edu/projects/glove/, zuletzt abgerufen am 28.12.2020.

[12] http://millionsongdataset.com/, zuletzt abgerufen am 28.12.2020. Auch als MSD abgekürzt.

Masterarbeit eine Orientierung bieten. Die Gegenüberstellung der Ergebnisse vorangegangener Arbeiten kann so beispielsweise in Form von Tabellen umgesetzt werden. Dazu sollen neben methodischen Aspekten wie der Wahl von FE- oder DL-Ansätzen auch die verwendeten Datensätze untersucht werden. Da sich die Datensätze als recht heterogen in der vorhandenen Literatur herausstellen, könnte dies einen nicht unerheblichen Einfluss auf die Ergebnisse der Emotionserkennung haben. Aber auch eine genaue Einordnung der verwendeten Emotionsmodelle scheint erforderlich, da es hier ebenfalls eine Vielzahl von unterschiedlichen Ansätzen zur Kategorisierung von Emotionen gibt. Ein Einfluss des verwendeten Emotionsmodells ist auf die Zuverlässigkeit der Emotionserkennung nicht auszuschließen und muss untersucht werden.

3.2 Emotionsklassifikator

Im zweiten Teil ist die Entwicklung eines eigenes Emotionsklassifikator geplant. Dazu soll eine neue, auf audio- und textbasierte Netzwerkarchitektur mithilfe von Deep-Learning-Methoden entwickelt werden. Hierfür scheint ein vielversprechender Ansatz die Kombination aus einem CNN für das Audiosignal und einem LSTM-Netzwerk für die Lyrics zu sein. Aber auch moderne Word Embeddings wie Facebooks *fastText* könnten in der Entwicklung der Netzwerkarchitektur eine interessante Rolle spielen. Die Zusammenführung der beiden Modalitäten innerhalb der Architektur könnte außerdem einen Schwerpunkt in der Entwicklung des Emotionsklassifkators darstellen. Zusätzlich könnten entsprechende FE- und DL-Ansätze aus vorangegangenen Arbeiten neu implementiert und mit dem eigenen Emotionsklassifikator verglichen werden. Die Vergleichbarkeit könnte hier durch die Verwendung eines gemeinsamen Datensatzes gewährleistet werden. Der Datensatz wird von dem Technologieunternehmen Cyanite[13] zur Verfügung gestellt.

4 Zeitplan

Der angestrebte Zeitplan findet sich in Abbildung 4.

<table>
<tr><td># Monat</td><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr>
<tr><td>Datum</td><td>Feb. 21</td><td>März 21</td><td>Apr. 21</td><td>Mai 21</td><td>Juni 21</td><td>Juli 21</td><td>Aug. 21</td></tr>
<tr><td></td><td>Exposé</td><td>An-
meldung</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td colspan="4">Recherche / Literatur / Datensatz / Übersichtsarbeit</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td colspan="3">Implementierung verschiedener Modelle / Architekturen</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="2">Vergleich</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="3">Auswertung / Schreiben</td></tr>
</table>

Abbildung 4: Zeitplan Masterarbeit.

[13] https://cyanite.ai/, zuletzt abgerufen am 19.01.2021.

Literatur

Bhattacharya, A. & Kadambari, K. (2018). A multimodal approach towards emotion recognition of music using audio and lyrical content. *arXiv preprint arXiv:1811.05760*.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C. & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96* (4), 668–696.

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J. & Moussallam, M. (2018). Music mood detection based on audio and lyrics with deep neural net. *Proc. of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 370—375.

Du, P., Li, X. & Gao, Y. (2020). Dynamic music emotion recognition based on cnn-bilstm. In *2020 ieee 5th information technology and mechatronics engineering conference (itoec)* (S. 1372–1376).

Eerola, T., Lartillot, O. & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Ismir* (S. 621–626).

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6* (3-4), 169–200.

Ekman, P. & Friesen, W. V. (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, *48* (2), 246–268.

Hu, X. & Downie, J. S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Ismir* (S. 67–72).

Hu, X. & Downie, J. S. (2010). When lyrics outperform audio for music mood classification: A feature analysis. In *Ismir* (S. 619–624).

Hu, X., Downie, J. S. & Ehmann, A. F. (2009). Lyric text mining in music mood classification. *American music*, *183* (5,049), 2–209.

Hu, Y., Chen, X. & Yang, D. (2009). Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Ismir* (S. 123–128).

Jeon, B., Kim, C., Kim, A., Kim, D., Park, J. & Ha, J. (2017). Music emotion recognition via end-to-end multimodal neural networks. In *Recsys posters.*

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proc. ismir* (Bd. 86, S. 937–952).

Laurier, C., Grivolla, J. & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *2008 seventh international conference on machine learning and applications* (S. 688–693).

Li, T. & Ogihara, M. (2003). Detecting emotion in music. In *4th International Symposium on Music Information Retrieval – ISMIR 2003* (S. 239—240). Johns Hopkins University.

Liu, T., Han, L., Ma, L. & Guo, D. (2018). Audio-based deep music emotion recognition. In *Aip conference proceedings* (Bd. 1967, S. 040021).

Liu, X., Chen, Q., Wu, X., Liu, Y. & Liu, Y. (2017). CNN based music emotion classification. *arXiv preprint arXiv:1704.05665*.

Lu, L., Liu, D. & Zhang, H.-J. (2005). Automatic mood detection and tracking of music

audio signals. *IEEE Transactions on audio, speech, and language processing*, *14* (1), 5–18.

Malheiro, R., Oliveira, H. G., Gomes, P. & Paiva, R. P. (2016). Keyword-based approach for lyrics emotion variation detection. In *Kdir* (S. 33–44).

Malheiro, R., Panda, R., Gomes, P. & Paiva, R. (2013). Music emotion recognition from lyrics: A comparative study. 6th International Workshop on Machine Learning and Music (MML13). Held in Conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPPKDD13).

Malheiro, R., Panda, R., Gomes, P. & Paiva, R. P. (2016). Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, *9* (2), 240–254.

Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D. & Jarina, R. (2017). Stacked convolutional and recurrent neural networks for music emotion recognition. *arXiv preprint arXiv:1706.02292*.

Panda, R. E. S. (2019). *Emotion-based analysis and classification of audio music* (Unveröffentlichte Dissertation). 00500:: Universidade de Coimbra.

Parisi, L., Francia, S., Olivastri, S. & Tavella, M. S. (2019). Exploiting synchronized lyrics and vocal features for music emotion detection. *arXiv preprint arXiv:1901.04831*.

Posner, J., Russell, J. A. & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17* (3), 715.

Pratt, C. C. (1952). Music as the language of emotion. The Library of Congress.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39* (6), 1161.

Schedl, M., Gómez Gutiérrez, E. & Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, *8* (2–3), 127–261.

Schuller, B., Dorfner, J. & Rigoll, G. (2010). Determination of nonprototypical valence and arousal in popular music: features and performances. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*, 1–19.

Van Zaanen, M. & Kanters, P. (2010). Automatic mood classification using tf* idf based on lyrics. In *Ismir* (S. 75–80).

Warriner, A. B., Kuperman, V. & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45* (4), 1191–1207.

Wieczorkowska, A., Synak, P. & Raś, Z. W. (2006). Multi-label classification of emotions in music. In *Intelligent Information Processing and Web Mining* (S. 307–315). Springer, Berlin, Heidelberg.

Yang, D. & Lee, W.-S. (2004). Disambiguating music emotion using software agents. In *Ismir* (Bd. 4, S. 218–223).

Yang, Y.-H. & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8* (3), 1–30.