Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

Genre-specific topics in song lyrics

Eine Projektarbeit von Niko Konzack, Benjamin Friedl und Lucas Zetzsche

Leipzig, Januar, 2024

Betreuender Hochschullehrer:
Dr. Andreas Niekler
Computional Humanities, Institut für Informatik

Inhalt

Einleitung [L]

Die Unterscheidung verschiedener Musikgenres anhand ihrer textlichen Merkmale stellt eine faszinierende Herausforderung dar, die nicht nur für Musikliebhaber, sondern auch für Datenwissenschaftler und Forscher von großem Interesse ist. Diese Forschungsfrage, ob typische Themen in den Liedtexten die Identifikation von Genres ermöglichen, ist von hoher Relevanz in musikbezogener Forschung und Datenanalyse. In diesem Bericht wird untersucht, ob und wie verschiedene Musikgenres anhand typischer Themen in ihren Liedtexten unterschieden werden können.

Beginnend mit dem Paper von (Rosebaugh und Shamir), das einen datenwissenschaftlichen Ansatz zur Vergleichbarkeit der Texte populärer Musikkünstler zeigt, wird deutlich, dass Textanalysen ein vielversprechender Ansatz sind, um die Vielfalt und Einzigartigkeit von Musikgenres zu erforschen. Obwohl diese Studie sich nicht explizit auf die Unterscheidung von Genres fokussiert, illustriert sie den Einsatz von Klassifikationsalgorithmen, um Texte anhand charakteristischer Merkmale zu unterscheiden. In ähnlicher Weise vertieft (Fell and Sportleder) diese Thematik, indem es sich mit der Klassifizierung von Genres basierend auf Textmerkmalen beschäftigt und diese mit menschlicher Wahrnehmung vergleicht. Die Ergebnisse solcher Studien legen nahe, dass die Analyse von Texten ein vielversprechender Ansatz ist, um die Vielfalt und Einzigartigkeit von Musikgenres zu erforschen.

Darüber hinaus bieten Untersuchungen wie von (Sterckx et al.) und (Lukic) einen Einblick in die Anwendung von Themenmodellen auf Liedtexte. Diese Arbeiten zeigen, dass Themenmodelle eine nützliche Methode sind, um charakteristische Themen in Liedtexten zu identifizieren und zu analysieren. Durch den Vergleich automatisch generierter Themen mit manuell erstellten Annotationen wird deutlich, dass eine höhere Datenqualität und -quantität zu präziseren Ergebnissen führen können.

Die eingehende Analyse von (Laoh) und (Sasaki et al.) beleuchtet die Herausforderungen bei der Modellierung von Themen in Texten, insbesondere in Bezug auf die indonesischen Liedtexte. Diese Studien zeigen, dass die sprachliche und kulturelle Vielfalt ein wichtiger Aspekt ist, der bei der Analyse von Musiktexten berücksichtigt werden muss.

Insgesamt legen diese vorangegangenen Arbeiten den Grundstein für die vorliegende Untersuchung, die sich mit der Frage befasst, ob bestimmte Themen typisch für bestimmte Musikgenres sind und zur automatisierten Genre-Klassifizierung verwendet werden können. Durch die Anwendung von Themenmodellen auf eine umfangreiche Textdatenbank verschiedener Genres wird untersucht, ob sich charakteristische Themen identifizieren lassen, die eine automatisierte Genre-Klassifizierung ermöglichen. Dabei wird ein Beitrag zur weiteren Entwicklung von Methoden geleistet, die es ermöglichen, Musikgenres auf der Grundlage ihrer textuellen Merkmale zu analysieren und zu unterscheiden.

Die vorangegangenen Studien bieten einen umfassenden Überblick über die Anwendung von Datenanalysen und Themenmodellen auf Musiktexte. Doch auch in anderen Forschungsbereichen wurden relevante Erkenntnisse erzielt, die einen wichtigen Beitrag zu

unserem Verständnis darüber liefern, wie Textanalysen zur Unterscheidung von Musikgenres beitragen können.

Ein vielversprechendes Feld ist die Stimmungsanalyse basierend auf Themen in Liedtexten, wie in (Bhattacharjee et al.) und (IEEE Staff and International Conference on Information Technology - the Next Generation IT Summit) dargestellt. Diese Arbeiten zeigen, dass Themenmodelle nicht nur dazu dienen können, Genreunterschiede zu identifizieren, sondern auch zur Erfassung und Klassifizierung von Stimmungen oder Themen in den Texten genutzt werden können. Die Ergebnisse solcher Studien tragen dazu bei, unser Verständnis darüber zu vertiefen, wie Textanalysen nicht nur zur musikalischen Genreerkennung, sondern auch zur Erfassung emotionaler oder thematischer Aspekte von Musiktexten beitragen können.

Des Weiteren liefern Studien wie (N. Rubin et al.) und (Kelechava) wichtige Einblicke in die Anwendung von Themenmodellen für die Klassifizierung von Musiktexten. Diese Arbeiten zeigen, dass die Verwendung von Themenverteilungen als Eingabe für Klassifikationsmodelle eine vielversprechende Methode ist, um verschiedene Genres oder Stimmungen in den Texten zu identifizieren und zu klassifizieren. Die Ergebnisse solcher Studien tragen dazu bei, unsere Methoden zur automatisierten Genre-Klassifizierung von Musiktexten zu verbessern und zu verfeinern.

Insgesamt verdeutlichen diese vorangegangenen Studien die Vielseitigkeit von Themenmodellen und Datenanalysen in der Musikforschung. Die vorliegende Untersuchung setzt an diesem Punkt an und trägt dazu bei, das Verständnis darüber zu vertiefen, wie verschiedene Musikgenres anhand typischer Themen in ihren Liedtexten unterschieden werden können. Durch die Anwendung von Themenmodellen auf eine umfangreiche Textdatenbank verschiedener Genres wird untersucht, ob sich charakteristische Themen identifizieren lassen, die eine automatisierte Genre-Klassifizierung ermöglichen. Dabei wird ein Beitrag zur weiteren Entwicklung von Methoden geleistet, die es ermöglichen, Musikgenres auf Grundlage von extrahierten Themen zu analysieren und zu unterscheiden.

Der im folgenden Bericht verwendete Code lässt sich unter https://github.com/Niko32/music-lyrics-analysis nachvollziehen.

Daten [N]

Die Wahl einer angemessenen Datengrundlage ist einer der wichtigsten Bestandteile für eine statistische Analyse. Dieses Kapitel befasst sich mit der Auswahl der Daten, ihrer Beschaffung für das Projekt und den Entscheidungen, die in diesem Kontext getroffen wurden.

Datenquelle [N]

Als Datengrundlage für das Projekt dient die API von Genius (ML Genius Holdings, LLC). Genius ist eine Online-Datenbank für Songtexte. Diese können hier sowohl von Künstlerinnen und Künstlern zu ihrer Musik selbst hochgeladen oder durch eine freiwillige Community bearbeitet werden. Zusätzlich zu den Texten selbst können Annotationen erstellt werden, welche einzelne Passagen genauer erklären und in einen Kontext einordnen können.
Für dieses Projekt ist vor allem ein Feature der Datenbank wichtig, welches es ermöglicht, Songs mit Tags zu versehen. Tags sollen es im Allgemeinen ermöglichen, einem Lied unterschiedliche Stichworte zuzuordnen, um diese semantisch gruppieren zu können. In erster Linie werden Tags von der Plattform und ihren Benutzern allerdings genutzt, um einem Lied ein oder mehrere Genres zuzuweisen. Neben der Nutzung der Tags für die Einordnung von Liedern in Genres, werden sie außerdem dazu genutzt, Lieder mit Sprachen und Herkunftsländern zu annotieren.
Genius pflegt eine erweiterbare Liste aller existierenden Tags (ML Genius Holdings, LLC). Diese unterscheidet die fünf *main tags* Country, Pop, Rap, R&B und Rock von über eintausend anderen *secondary tags*. Die *secondary tags* sind oftmals zusammengesetzt aus den *main tags* (z.B. “Pop Rock”). Zusätzlich gibt es aber auch gänzlich andere Genrebezeichnungen, welche sich schwer oder gar nicht auf die *main tags* zurückführen lassen (z.B. “Reggae”) und Kombinationen aus diesen mit anderen Tags (z.B. “Reggae Rock”) oder feiner definierte Subgenres (z.B. “French Reggae”).

Auswahl der Genres [N]

Ziel des Projektes ist es, Topics innerhalb unterschiedlicher Genres zu untersuchen. Voraussetzung hierfür ist es, eine sinnvolle Menge an Genres zu definieren, die betrachtet werden soll. Die Untersuchung der mehr als eintausend *secondary tags* kommt nicht infrage, da die meisten dieser Tags nur sehr spärlich besetzt sind und somit keine gute Datengrundlage bieten. Ein besserer Ansatz wäre die Verwendung der fünf *main tags.* Da diese allein nach Auffassung der Autoren jedoch keine breite Menge an Musikrichtungen abdecken, ergibt es Sinn, die Auswahl noch zu erweitern.
Mit dem Free Music Archive (FMA) (Defferrard et al.) wird 2016 eine Musikdatenbank ins Leben gerufen, das mit Audio-Samples von über 100.000 Liedern eine offene Grundlage für Forschung im Bereich Music Information Retrieval (MRI) bereitstellen soll. Neben anderen wichtigen Beiträgen für dieses Gebiet stellt das FMA auch eine hierarchische Definition von Musikgenres. Insgesamt werden 161 Genres benannt, von denen 16 Genres als top-level Genres bezeichnet werden. Diese 16 top-level Genres sind *International*, *Blues*, *Jazz*, *Novelty*, *Historic*, *Country*, *Pop*, *Instrumental*, *Rock*, *Soul-RNB*, *Spoken*, *Experimental*, *Folk*, *Classical*, *Electronic* und *Hip-Hop*. Diese Liste enthält neben den *main tags* von Genius noch

einige weitere Genres und ermöglicht dadurch eine viel breitere Abdeckung der gesamten Musiklandschaft.
Von den 16 top-level Genres können jedoch nicht alle betrachtet werden. Das Genre *International* beinhaltet beispielsweise viele Subgenres, die für jeweils eine bestimmte Region der Welt und ihrer Musikkultur stehen. Das top-level Genre selbst bietet in diesem Fall selbst kaum eine eigene Semantik. Stattdessen dient es als Container für ähnliche Subgenres. Das Genre *Instrumental* ist für das Thema des Projektes ungeeignet, da bei Liedern mit diesem Genre kaum Lyrics zu erwarten sind. Die Genres *Novelty*, *Historic* und *Spoken* können zudem nicht betrachtet werden, da diese in der Liste der Tags von Genius so nicht vorhanden sind. Übrig bleiben für die Analyse die fünf *main tags* von Genius und die Genres *Blues*, *Jazz*, *Experimental*, *Folk*, *Classical* und *Electronic*. Damit werden insgesamt elf verschiedene Genres betrachtet.

Erstellung eines Datensatzes [N]

Für die Erstellung eines Datensatzes mit den genannten Genres wird die API von Genius (ML Genius Holdings, LLC) genutzt. Die Bibliothek *lyricsgenius* (Miller) implementiert die Authentifizierung und Abfrage der dort verfügbaren Endpunkte in Python. Einer dieser verfügbaren Endpunkte ermöglicht die Abfrage der populärsten Lieder zu einem gegebenen Tag. Da diese Abfrage für jeden Tag nur bis zu 1000 Lieder ausgibt, ist es sinnvoll, einen Datensatz mit 1000 Songs pro Tag zu erstellen. Eine solche Liste von 1000 Liedern pro Genre befindet sich im Git unter *data/multilabel/songs* im YAML-Format.
Ein Problem, das mit diesem Vorgehen auftritt, ist, dass ein Lied in mehreren dieser Listen auftauchen kann. Der so entstandene Datensatz eignet sich also für eine *multilabel* Klassifizierung, weniger aber für ein *multiclass* Modell, das eigentlich angestrebt wird. Es wird also ein weiterer Datensatz benötigt, der jedem Lied nur einen einzigen Tag zuweist. Um diesen zu erstellen, muss ein Weg gefunden werden, aus einer vorhandenen Liste von Tags für jedes Lied einen einzigen Tag zu finden, der auf den gegebenen Song am meisten zutrifft. Obwohl es für die Reihenfolge der Tags in der Liste keine Vorschriften von Genius gibt, fällt bei der Untersuchung einzelner Beispiellieder auf, dass die relevantesten Tags zuerst genannt werden. Weiter hinten in der Liste tauchen dann zunehmend Tags auf, die nach subjektiver Einschätzung weniger auf den Song zutreffen. Außerdem werden einem Lied oftmals zuerst allgemeine Beschreibungen hinzugefügt (z.B. “Pop”) bevor spezifische Bezeichnungen vergeben werden (z.B. “Elektro-Pop”). Wir gehen davon aus, dass bei der Vergabe der Tags zuerst die Tags genannt werden, die für den Autor oder die Autorin des Eintrags am offensichtlichsten erscheinen. Unter dieser Annahme lässt sich ein einziger mutmaßlich relevantester Tag für einen Song bestimmen. Dieser wird definiert als der erste Tag aus der Liste unserer elf betrachteten Genres, welcher in der Liste von Tags eines Songs auftritt.
Im so entstandenen Datensatz gibt es keine Duplikate mehr in den Songlisten der einzelnen Genres. Infolgedessen sinkt auch die Anzahl der Lieder, welchen einem einzelnen Genre zugeordnet sind. Das trifft vor allem auf die Genres zu, die Genius nicht als *main tag* definiert, da die meisten Songs zuerst mit beispielsweise “Pop” annotiert sind bevor eine bezeichnung wie “Electronic” vergeben wird. So sind von den 1000 Songs, welche für den *multilabel* Fall zum Genre “Experimental” zugeordnet wurden, nur 57 in der *multiclass* Liste übrig. Um die Anzahl der Lieder in den kleineren Genres aufzufüllen, mussten deswegen weitere Lieder von der API abgerufen werden. Da die Liste der 1000 Songs pro Genre erschöpft ist, wurden diese über zufällig gezogene Song-IDs angefragt. Wenn ein zufälliger

Song eines der seltenen Genres an erster Stelle aufwies, wurde es dem Datensatz hinzugefügt. Auf diese Weise konnte für die *multiclass* Aufgabe eine Liste von 150 Songs pro Genre zusammengestellt werden. Eine separate Liste mit 850 Songs pro Genre wurde für die ausreichend vorhandenen *main tags* erstellt. Um zu vermeiden, dass eine Klasse im Datensatz stärker repräsentiert wird als die anderen, richten sich die Anzahlen nach dem Genre mit den wenigsten verfügbaren Liedern. Aus den anderen Genres wurde die gleiche Anzahl an Liedern zufällig gezogen. Die beiden Listen sind zu finden unter *data/pruned/all_tags_songs.csv* und *data/pruned/main_tags_songs.csv* mit entsprechenden Lyrics für jede Song-ID unter *data/pruned/lyrics*.

Methoden [B]

Ziel ist es, herauszufinden, ob sich Musikgenres anhand der Themen von Song-Lyrics trennen lassen. Dazu müssen zuerst automatisiert Themen gefunden werden. Dies geschieht durch die sogenannte Topic Modellierung.

Topic Modellierung [B]

Topic Modelle sind statistische Modelle zur Erkennung sogenannter „Topics“ in großen Textkorpora. Die Topics werden hierbei durch Muster ko-existierender Wörter repräsentiert und geben Einblick in versteckte semantische Strukturen der Texte. Dies ist insbesondere in der Analyse großer Textkorpora von Bedeutung, wo klassische, nicht-automatisierte Verfahren aufgrund der reinen Menge an Texten an ihre Grenzen stoßen. In unserem Fall sollen so Themen eines Korpus an Lyrics und eine Zuordnung jeder Lyric zu diesen Themen gefunden werden.

Eine der populärsten und meistgenutzten Topic Modelle ist die Latent Dirichlet Allocation (LDA) (Blei). Sie gehört zu den probabilistischen Topic Modellen und arbeitet auf Basis von Dirichlet Verteilungen.

Terminologie: (Chauhan and Chauhan)

- Korpus: Eine Menge von M Dokumenten $D = \{W_1, \dots, W_M\}$
- Dokument: Eine Menge von N_m Wörtern $W = \{w_1, \dots, w_{N_m}\}$, für $m \in \{1,\dots, M\}$
- Wort: Zentrale Einheit des Textes, repräsentiert durch $w \in \{1,\dots, V\}$
- Topic: Wahrscheinlichkeitsverteilung über die Menge von V Wörtern, beschrieben als $z \in \{1,\dots, K\}$.
- α: Parameter für den Dirichlet-Prior für die Dokument-Topic Verteilung.
- η: Parameter für den Dirichlet-Prior für die Topic-Word Verteilung.
- Θ_m: Topic Verteilung für Dokument m.
- ϕ_k: Wort Verteilung für Topic k.

Generativer Prozess: (Chauhan and Chauhan)

Sei $D = \{W_1, \dots, W_M\}$ ein Korpus. Für jedes $m \in \{1,\dots, M\}$:

- Wähle eine K-dimensionalen Topic-Verteilung Θ_m über $p(\Theta_m|\boldsymbol{\alpha}) = Dir_K(\boldsymbol{\alpha})$
- Wähle eine V-dimensionale Wort-Verteilung ϕ_k über $p(\phi_k|\eta) = Dir_V(\eta)$
- Für jedes Wort mit index $n \in \{1, \ldots, N_m\}$:
    - Wähle ein Topic $z_n \in \{1,\ldots, K\}$ über die Verteilung $p(z_n = k|\Theta_m)$
    - Für dieses Topic z_n, wähle ein Wort w_n über $p(w_n|z_n = j, \eta)$

Dies definiert die folgende gemeinsame Verteilung über den Topic-Mix Θ, die Menge aller Topics z und die Menge aller Wörter w:

$$p(\Theta, z, w|\boldsymbol{\alpha}, \eta) = \prod_{k=1}^{K} p(\phi_k|\eta) \prod_{m=1}^{M} p(\Theta_m|\boldsymbol{\alpha}) \prod_{n=1}^{N_m} p(z_n|\Theta_m) p(w_n|z_n, \eta)$$

Für dieses Modell sind die Menge aller Topics K und die Verteilungsparameter $\boldsymbol{\alpha}$ und η festzulegen. Um sicherzustellen, dass unser Topic Modell in der Lage ist, charakteristische Themen der Lyrics zu finden und schlechte Ergebnisse nicht aufgrund eines mangelhaften Modells auftreten, müssen durch einen strukturierten Prozess Werte für diese Parameter ausgesucht werden. Details hierzu können unter “Ergebnisse: Modell Selektion” gefunden werden.

Evaluierungsmethoden [B] [L]

Wurde ein repräsentatives Topic Modell ausgewählt, gilt es dessen gefundenen Topics zu interpretieren. Uns interessiert dabei spezifisch die Trennbarkeit von Genres. Die Evaluation erfolgt dabei auf zwei Arten: die direkte, menschliche Interpretation der Topics und die Evaluation mit Hilfe eines Klassifikators.

Evaluierung: Menschliche Interpretation [L]

Bei der Evaluierung durch menschliche Interpretation werden verschiedene Methoden angewandt, um die gefundenen Topics zu analysieren und ihre Relevanz für die Unterscheidung von Genres zu bewerten. Eine wichtige Aufgabe besteht darin, die signifikantesten Topics des gesamten Modells zu identifizieren und zu analysieren. Dazu werden die gefundenen Topics auf ihre Kohärenz und Interpretierbarkeit geprüft, um festzustellen, ob sie relevante Aspekte der Liedtexte abbilden. Zudem werden aus einer bestimmten Anzahl an Topics jeweils die aussagekräftigsten Terme herausgefiltert und evaluiert, um ihre Bedeutung für die Genreunterscheidung zu untersuchen.

Ein weiterer Schritt besteht darin, die Topics nach Genres zu untersuchen. Dabei werden die gefundenen Topics für ausgewählte Genre analysiert, um festzustellen, ob sich charakteristische Themen identifizieren lassen, die typisch für dieses Genre sind. Dies ermöglicht es, die Trennbarkeit der Genres auf der Grundlage der gefundenen Topics zu bewerten und Rückschlüsse auf die Eignung des Topic-Modells für die automatisierte Genre-Klassifizierung zu ziehen.

Menschliche Interpretation bietet einen unschätzbaren Vorteil gegenüber rein algorithmischen Ansätzen: Menschen sind immer noch in der Lage, komplexe Zusammenhänge besser zu erfassen und zu bewerten, selbst wenn diese nicht so leicht skalierbar sind wie maschinelle Verfahren. Ein weiterer entscheidender Aspekt ist die Fähigkeit des Menschen, Fehler oder irrelevante Wörter sofort zu erkennen, was bei der automatisierten Auswertung durch zum Beispiel einen Klassifikator nicht unbedingt der Fall ist und es leicht zu verfälschten Ergebnissen führen könnte. Daher ist die menschliche Interpretation ein unverzichtbares Element bei der Bewertung und Validierung von Topic-Modellen in der Textanalyse.

Die Evaluierung durch menschliche Interpretation bietet somit eine wichtige qualitative Bewertung der Trennbarkeit der Genres auf der Grundlage der gefundenen Topics. Sie ermöglicht es, die Ergebnisse des Topic-Modells kritisch zu hinterfragen und zu validieren. Dennoch ist es wichtig, diese Evaluierungsmethode mit numerischen Evaluierungsmethoden, wie der Evaluation mit Hilfe eines Klassifikators, zu kombinieren, um ein umfassendes Bild der Trennbarkeit der Genres zu erhalten.

Evaluierung: Klassifikator [B]

Durch das Topic Modell ist es möglich für jede Song-Lyric die Anteile unterschiedlicher Topics herauszufinden. Sind diese Anteile für Lyrics unterschiedlicher Genres gleich, so kann auch ein Klassifikator mit diesen Anteilen als Features nicht mehr zwischen den Genres unterscheiden. Je unterschiedlicher jedoch diese Anteile sind und je isolierter bestimmte Topics bestimmten Genres zuzuordnen sind, desto besser kann auch ein Klassifikator zwischen den Genres unterscheiden und desto höher ist dessen Performance. Das ist die grundlegende Idee hinter dieser numerischen Evaluierungsmethode zur Quantifizierung der Trennbarkeit der Genres. Die Bewertung dieser Ergebnisse muss allerdings auch im Kontext der Schwierigkeit der Aufgabe, also der Klassifizierung von Genres aufgrund von Song-Lyrics, erfolgen. Dazu müssen bewährte Text-Klassifikations-Modelle implementiert und als Performance-Baselines herangezogen werden. Diese Methode ersetzt allerdings nicht die menschliche Interpretation. Inwiefern unterschiedliche Topics auch tatsächlich unterschiedliche Themen behandeln und nicht nur genre-spezifische sprachliche Unterschiede aufweisen, kann so nicht bewertet werden. Auch hängt die Performance nicht nur von der tatsächlichen Trennbarkeit der Genres, sondern auch von dem Klassifikationsmodell und der Menge an zur Verfügung stehenden Daten ab.

Ergebnisse [B] [L]

Experimentelles Setup [B]

Die Experimente wurden in Python implementiert. Für das Topic Modelling wurde die gensim Bibliothek genutzt, für die Klassifikation sci-kit learn (Radim and Sojka, Pedregosa et al.). Es werden die oben beschriebenen Daten benutzt. Sie werden in Trainings- und Testdaten aufgeteilt. Die Trainingsdaten werden sowohl zum Training des Topic Modells als auch zum

Training des Klassifikators benutzt. Dies ist möglich, da das Training des Topic Modells ein unsupervised Prozess ist, in den keine Genre-Informationen mit hineinfließen. Die Testdaten dienen schließlich der Evaluierung des Klassifikators.

Vorverarbeitung [B]

Um die Lyrics für das Topic Model handhabbar zu machen und um bessere Ergebnisse zu erzielen, durchlaufen die Lyrics mehrere Vorverarbeitungsschritte. Diese Schritte können hier grob in zwei Teile aufgeteilt werden: Das Säubern der Lyrics und das Erstellen von Wörterbüchern und Korpussen, die für das Topic Model benötigt werden.

Säuberung der Lyrics [B]

Die Säuberung der Lyrics umfasst das Weglassen nicht-englischer Lyrics, das Entfernen von Satzzeichen und das Umformen jeglicher Wörter in Kleinschreibung. Des Weiteren sind in den Lyrics noch Meta-Informationen als Zeilen der Form [Verse 1], [Chorus], [Bridge], etc. enthalten. Diese werden mittels einer Regular-Expression entfernt. Außerdem werden Zeilenumbrüche und anschließend Folgen von Leerzeichen entfernt. Details können dem GitHub Code entnommen werden. Außerdem ist zu beachten, dass Stoppwörter und zu häufig-/niedrig-frequente Wörter bei der Erstellung der Corpora entfernt werden. Als Evaluation dieses Schrittes dient eine Wordcloud (siehe Fig. 1). Zum einen ist hier zu erkennen, dass keine unerwarteten Wörter auftauchen. In ersten Versuchen waren hier zum Beispiel noch Wörter nicht-englischer Sprachen sowie Wörter, die nicht richtig kodiert wurden, enthalten. Zum anderen können hier häufige Wörter identifiziert werden, die nicht unserer Forschungsfrage dienlich sind. Beispiele hierfür sind Wörter wie „oh“, „la“, „yeah“, die wenig semantische Bedeutung haben und folglich keine Themen charakterisieren können, anhand derer wir Genres trennen wollen. Sie werden als Stoppwörter im nächsten Schritt entfernt.

Die große Frage hier ist, ob Verben wie „know“, „might“, „make“, etc., die diese Wordcloud dominieren, in der Lage sind solche Themen zu charakterisieren und wenn ja, ob sie zu einer besseren Trennung der Genres führen. Sie könnten durch ihre allgemeine Häufigkeit jegliche Topics dominieren und folglich den Platz charakteristischerer Wörter einnehmen. Sie könnten allerdings auch informativ sein. Verben wie „stay“, „leave“ deuten auf Liebeskummer hin. Dass Verben wie „could“, „would“ anstelle von „can“, „will“ verwendet werden, könnte auf einen Unterschied in der von der Sänger*in wahrgenommenen Wahrscheinlichkeit und Distanz bestimmter Ereignisse hindeuten und charakteristisch für spezifische Genres sein. Dies ist im Voraus schwierig zu sagen, daher werden zwei Corpora und Dictionaries erstellt. Eines mit diesen Verben und eines ohne.

Fig. 1: Wordcloud nach Säuberung der Lyrics

Erstellen von Wörterbüchern und Korpora

Am Anfang dieses Prozesses steht das Entfernen von Stoppwörtern. Diese umfassen Standard-Stoppwörter der englischen Sprache des Pakets „nltk“ und weitere in der Wordcloud auffällige Wörter wie „oh“ oder „la“. Eine von zwei Versionen dieser Stoppwörter enthält außerdem weitere Verben (siehe oben). Es entstehen folglich zwei Versionen jedes Korpus und Wörterbuchs. Nach Entfernung der Stoppwörter werden mit Hilfe der Phrases-Funktion des Pakets „gensim“ Bigramme erstellt. Bei der Erstellung des Wörterbuchs, wird jedem Wort/Bigramm eine ID zugewiesen und zu seltene/häufige Wörter, bzw. Bigramme entfernt. Hier ist zu beachten, dass wir aufgrund der späteren Implementierung eines Klassifikators mit Trainings- und Testdaten zu tun haben. Da für die Testdaten später das Topic Model benutzt wird, welches auf den Trainingsdaten trainiert wurde, muss für den Testkorpus dasselbe Wörterbuch benutzt werden. Der Korpus wird anschließend als Liste jedes Dokuments als Bag of Words erstellt.

Am Ende des Vorverarbeitungs-Prozesses stehen folglich Wörterbuch, Trainings- und Testkorpus einmal mit und einmal ohne entfernte Verben zur Verfügung.

Topic Modell Selektion [B]

Um Lyrics anhand ihrer Themen durch Topic Modelling zu bewerten, muss sichergestellt werden, dass das ausgesuchte Topic Modell in der Lage ist charakteristische Themen zu finden und potenziell schlechte Ergebnisse nicht aufgrund eines mangelhaften Modells entstehen. Das Topic Modell wird daher durch einen strukturierten Prozess ausgesucht, der in mehrere Schritte unterteilt ist. Aufgrund des Aufwands dieses Prozesses wird dieser lediglich für den Trainingskorpus mit Verben durchgeführt.

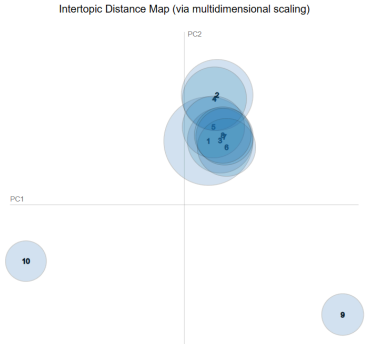
Intuition: Anzahl an Topics [B]

Zuerst wird ein Hierarchical Dirichlet Process (HDP) Modell trainiert (Teh et al.). Dieses hat die Besonderheit, dass es automatisch eine gute Anzahl an Topics findet und wird hier als Intuition zur Festlegung der Anzahl an Topics für LDA benutzt. Über eine Bandbreite an

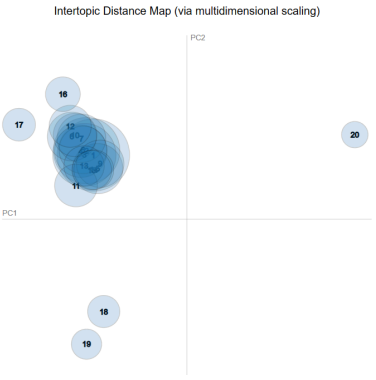
Werten α wurden 150 Topics festgelegt. Die herausgefundenen Topics waren zudem für geringere Werte von α auch interpretierbar und dennoch unterschiedlich. Beides sind wichtige Voraussetzungen, um später anhand der Themen Genres trennen zu können.

Testen verschiedener Anzahlen an Topics und Werten alpha [B]

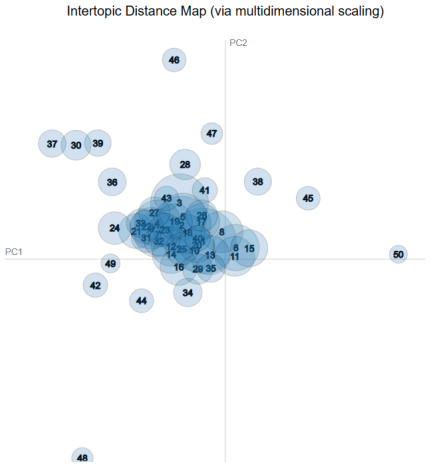
Im Folgenden werden nun Topic Modelle für Anzahlen an Topics $K \in \{10, 20, 50, 100, 150\}$ getestet. Dabei werden für jede Anzahl an Topics Werte für $\alpha \in \{0.0625, 0.125, 0.25, 0.5, 1.0, 2.0\}$ evaluiert und das Beste herausgesucht. η wird immer auf 1 / K gesetzt. Die Topics des ausgesuchten Modells werden anschließend im zwei-dimensionalen Raum mit Hilfe einer Intertopic Distance Map (IDM), implementiert durch pyLDAvis, visualisiert und kurz interpretiert.



10 Topics: Die entstandenen Topics sind ziemlich ähnlich, sowohl innerhalb eines Topic Modells als auch zwischen den Topic Modellen für unterschiedliche Werte von α. Die meisten beschreiben eine Liebesbeziehung, z.B. baby/say, baby/tell/me oder come/back. Die Log-Perplexitäten sind auch sehr ähnlich unabhängig von α. Für kleinere Werte alpha sind die Topics allerdings etwas unterschiedlicher, gewählt wurde also α=0.0625. Die IDM zeigt auch ein großes Cluster rund um Liebesbeziehungen. Die beiden Ausreißer-Topics sind schwierig zu interpretieren.

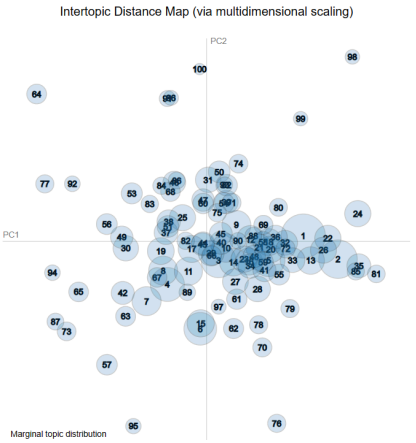


20 Topics: Ähnlich wie bei 10 Topics sind sich alle Topics recht ähnlich, sowohl innerhalb jedes Topic Modells als auch zwischen Topic Modellen verschiedener Werte von α. Für Werte von α in [1.0,2.0] bestehen die Topics allerdings maßgeblich aus den Wörtern „baby“, „never“, „time“. Für Werte von α in [0.0625,0.125,0.25,0.5] sind die Topics ähnlich wenig divers. Da die Log-Perplexität für α=0.0625 am geringsten ist, wird dieses Modell ausgesucht. Die IDM zeigt wieder ein großes Cluster rund um Liebesbeziehungen mit vereinzelten Ausreißern. Die Ausreißer hier sind allerdings interpretierbar. Zum Beispiel behandelt Topic 18 (baby/come/time/come_back) den Wunsch jemanden zurückzubekommen, oder Topic 19 (let/think/life/time) ein Nachdenken über das Leben.

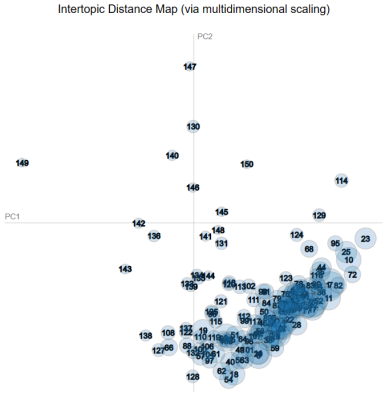


50 Topics: Hier lassen sich erstmals größere Unterschiede zwischen den Topics für unterschiedliche Werte von α beobachten. Während für alpha in [1.0,2.0] die Topics wieder recht generisch Liebesbeziehungen beschreiben, ergeben sich für kleinere Werte von α Themen abseits dieses Themenkomplexes. Da die Topics für α=0.125 am interpretierbarsten erschienen, wurde dieses Modell ausgewählt. Die IDM bekräftigt die wahrgenommene Distanz zwischen den Topics. Es gibt zwar ein großes Cluster, dieses ist allerdings wesentlich verteilter. Außerdem lassen sich unterschiedliche Quadranten dieses Graphen unterschiedlichen Themen zuordnen. Der Quadrant unten links ist mehr dem Thema Leben und weniger dem Thema Liebe zuzuordnen. Wörter wie

„world“, „generation“, „think“ kommen häufig in den Topics vor. Die rechte Hälfte charakterisiert sich besonders durch unterschiedliche Sprache wie „shit“ oder „fuck“ und deutet auf durch Hip-Hop dominierte Topics hin.



100 Topics: Ähnlich zu 50 Topics, gibt es hier größere Unterschiede zwischen den Topics für unterschiedliche Werte von α. Für α in [1.0,2.0] ist fast jedes Topic dominiert von „go“, „baby“, „never“, „time“. Für die restlichen α's ergeben sich unterschiedliche Topics. Die meisten der 20 relevantesten Topics drehen sich zwar immer noch um Liebesbeziehungen, allerdings von unterschiedlichen Perspektiven. Die Log-Perplexitäten sind ziemlich ähnlich, am besten zu interpretieren sind allerdings die Ergebnisse für α=0.125. Die IDM zeigt, wie gewünscht, stark verteilte Topics. Es gibt auch Zusammenhänge zwischen der Position in der IDM und thematischen Einordnungen. Der Quadrant unten links behandelt beispielsweise viel „needing“, „girl“ und „never“, „let“, „go“. Der Quadrant unten rechts ist sprachlich eher Hip-Hop dominiert.



150 Topics: Ähnlich wie bei 50 und 100 Topics sind auch hier die Topics für α in [1.0,2.0] recht generisch, während sich für die restlichen Werte von α diverse Topics ergeben. Die wahrgenommene Interpretierbarkeit der Topics unterscheidet sich hier allerdings nicht. Da für α=0.0625 die Log-Perplexität am geringsten ausfällt, wird dieses Modell ausgewählt. Die IDM zeigt weniger verteilte Topics im Gegensatz zu 50 und 100 Topics. Vielmehr gibt es ein großes Cluster im Quadranten unten rechts. Dennoch sind die Topics innerhalb dieses Clusters divers. Sie behandeln nicht nur Liebesbeziehungen, sondern auch Alkoholismus (Topic 101), oder Träume über Amerika (Topic 80) inmitten dieses Clusters.

Der Verlauf der Log-Perplexitäten der ausgewählten Modelle über die Anzahl in Topics ist in Fig. 2 erkennbar. Aufgrund eines Ellbogens bei 100 Topics und der starken Verteilung und Interpretierbarkeit dieser Topics wird dieses Modell ausgewählt.

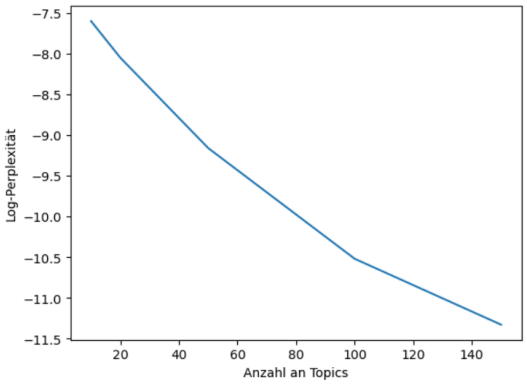


Fig. 2: Verlauf der Perplexitäten über Anzahl an Topics

Menschliche Evaluation des Topic Modells [L]

Im Folgenden geschieht die menschliche Evaluation zum einen auf dem Topic Modell mit Verben, sowie dem Topic Modell ohne Verben. Hierfür werden zunächst die Topics des gesamten Modells näher angeschaut und wie diese in Beziehung zueinander stehen. Des Weiteren werden die 10 signifikantesten Topics und deren Terme pro Genre evaluiert und zum Schluss werden Fehlerquellen betrachtet, die ein Klassifikator nicht ohne weiteres erkennen würde.

Topic Modell mit Verben [L]

Die Evaluation der 100 Topics des gesamten Modells zeigt auf, dass viele Topics sich sehr ähneln und dass viele gleiche Terme genutzt werden (siehe Fig. 3 und 4).

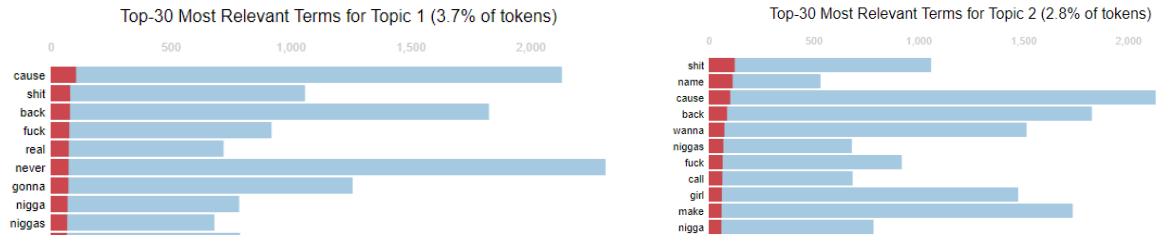


Fig. 3: Vergleich der zwei größten Topics die auf das Rap-Genre hinweisen

Eine Unterscheidung der Topics fällt besonders schwer aufgrund von Wörtern wie “never”, “let”, “go”, “come” oder auch “baby”. Diese Terme können im Zusammenhang mit anderen Worten zwar Sinn ergeben, jedoch helfen sie bei der Unterscheidung von Genres sehr wenig. Ähnlich verhalten sich Slangwörter wie “wanna”, “gimme, “gotta” oder "gonna", die Hinweise auf die Genre Richtung geben, jedoch noch nicht aussagekräftig genug sind. Eine erste wichtige Schlussfolgerung ist daher, dass es viele Topics geben kann, diese aber nur ein Thema behandeln.

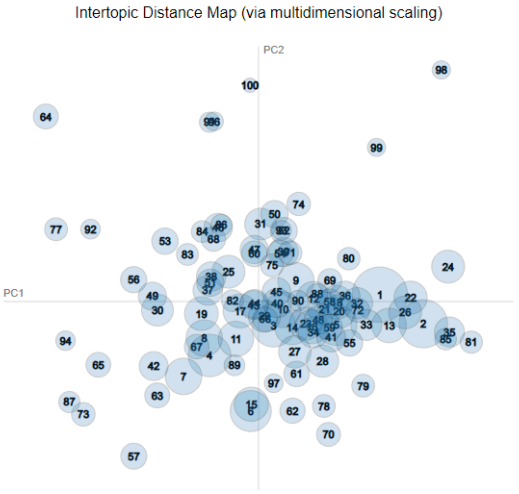


Fig. 5 Verteilung der Topics mit Verben

In der Gesamtverteilung lassen sich so grobe Themencluster erkennen. Diese decken sich auch mit den Clustern, die man erhält, wenn man sich die Topics pro Genre anschaut.

Die Untersuchung der einzelnen Topics pro Genre ermöglicht eine detailliertere Analyse der behandlungsrelevanten Themen und Begriffe. Dabei werden die charakteristischsten Begriffe in den jeweiligen Topics betrachtet, um typische Themenbereiche für jedes Genre zu identifizieren.

Im Pop-Genre werden verschiedene Themenbereiche behandelt, wobei Liebe und Beziehungen eine zentrale Rolle spielen. Begriffe wie "woman", "heart" und "love" weisen auf romantische Beziehungen hin, während Wörter wie "come", "night" und "baby" auf romantische Treffen oder Intimität hinweisen. Neben diesen romantischen Aspekten behandeln Pop-Songs auch Themen der Einsamkeit. Begriffe wie "alone", "forever" und "waiting" geben Einsamkeit oder Verlust wieder, während Wörter wie "night", "dark" und "lonely" emotionale Tiefen oder Melancholie reflektieren. Zusätzlich finden sich in den Pop-Texten Elemente der Natur wie "sun", "star" und "dance", die eine positive Atmosphäre und Freude vermitteln. Diese relativ allgemeinen Themen, die das Pop Genre auch ausmachen, kommen allerdings auch in anderen Genren auf und machen es daher besonders schwer, hier eine überzeugte Zuordnung zu machen.

Im R&B-Genre werden Liebesbeziehungen oft auch aus einer traurigen Perspektive betrachtet. Begriffe wie "girl", "wanna" und "tell" zeigen emotionale Bindungen oder Sehnsucht an, während Wörter wie "heart", "never" und "tears" Herzschmerz oder Verlust reflektieren. Neben diesen emotional belastenden Themen behandeln R&B-Songs auch allgemein melancholische Themen. Viele Begriffe wie "sad", "cry" und "pain" spiegeln emotionale Schmerzen oder Traurigkeit wider. Die Zuordnung von Lyrics zum R&B fällt aufgrund des afroamerikanischen Einflusses und den damit verbundenen Termen der Black-Community etwas leichter als beim Pop Genre, allerdings ist es wie beim Pop oftmals nur ein Raten.

Im Rock-Genre sind verschiedene Themenbereiche präsent. Eine häufig auftretende Thematik ist die Suche nach einem besseren Leben und die Veränderung. Wörter wie "keep", "go" und "walk" könnten auf das Streben nach Verbesserung oder Freiheit hinweisen, während Begriffe wie "never", "learn" und "home" den Wunsch nach Stabilität oder Veränderung reflektieren. Neben diesen Themen behandeln Rock-Songs auch Konzepte der Liebe und deren Beständigkeit. Begriffe wie "never", "baby" und "ever" betonen die Beständigkeit von Gefühlen oder Bindungen, während "eyes", "always" und "love" eine emotionale Verbundenheit oder Treue vermitteln.

Im Country-Genre werden verschiedene Themen angesprochen, wobei häufig ein Bezug zum Glauben zu finden ist. Begriffe wie "lord", "time" und "new" könnten spirituelle Themen oder Hoffnung auf Veränderung repräsentieren, während "meet", "man" und "world" auf zwischenmenschliche Beziehungen oder das Landleben hinweisen. Des Weiteren behandeln Country-Songs oft Erinnerungen an vergangene Liebe. Wörter wie "old", "let" und "still" zeigen eine nostalgische Stimmung oder die Sehnsucht nach Vergangenem, während "heart", "gone" und "darling" Gefühle der Verlusttrauer oder der Veränderung widerspiegeln.

Im Rap-Genre zeigt sich eine Vielfalt an Themen und Sprachgebrauch. Der Wortschatz und Ausdruck unterscheidet sich oft deutlich von anderen Genres, wobei Begriffe wie "bitch", "ya", "yo" und "motherfucker" eine direktere und oft explizitere Sprache im Rap widerspiegeln. Des Weiteren thematisieren Rap-Songs häufig Liebe, wobei Wörter wie "love", "together" und "roll" auf emotionale Bindungen oder Lebensstile hinweisen, während "real", "feel" und "hate" die Authentizität und Echtheit der Erfahrungen reflektieren. Zudem finden sich religiöse Bezüge in den Texten, wie beispielsweise "jesus", "god" und "lord", die spirituelle Einflüsse oder Konzepte repräsentieren. Darüber hinaus behandeln Rap-Songs oft Erfolg und Lebensstil, wobei Wörter wie "rap", "high" und "money" den Aufstieg, Reichtum oder den Lebensstil im Rap-Genre betonen.

Diese Analyse der Topics pro Genre verdeutlicht die vielfältigen Themen und Ausdrucksformen in den verschiedenen Genres sowie die charakteristischen Begriffe, die jedem Genre zugeordnet werden können. Dabei zeigen sich auch viele Überschneidungen und Variationen innerhalb der Genres, was die Komplexität und Vielschichtigkeit der Musiktexte unterstreicht, aber auch die Schwierigkeit der Zuordnung verdeutlicht.

Die menschliche Evaluation bietet eine entscheidende Ergänzung zur rein algorithmischen Analyse von Topic-Modellen, da sie bestimmte Aspekte der Daten und Ergebnisse aufdecken kann, die für Maschinen schwer zu erfassen sind. Ein Beispiel dafür ist die Fähigkeit des Menschen, kontextuelle Bedeutungen und metaphorische Ausdrücke zu erkennen. Während ein Algorithmus möglicherweise nur die Wörter selbst betrachtet, kann ein Mensch tiefer in den Kontext eintauchen und verstehen, wie bestimmte Begriffe in verschiedenen Kontexten verwendet werden. Zum Beispiel könnte ein Algorithmus das Wort "fly" in Kombination mit “high” oder “wings” nicht sicher einordnen, während ein Mensch erkennen könnte, dass es metaphorisch für den Abschied zu einem Verstorbenen bei “fly” plus “high” steht und dies oft in Rap Texten verwendet wird. “Fly” und “wings” in der Kombination könnte dagegen für die Unabhängigkeit und Freiheit stehen, die eher auf Pop und Rock hinweisen. Ein weiteres Beispiel ist das Wort “stick” im Zusammenhang von Wortgruppen wie “gun”, “chasing” und “rap” wird schnell klar das damit eine Pistole gemeint ist und nicht das kleben oder beiseite stehen wie in der Kombination von “friends” und “togehter”.

Eine damit verbundene Fehlerquelle sind Wörter, die besonders ausschlaggebend für die Zuordnung zu einem Genre sind, aber nicht vom Algorithmus als aussagekräftig genug gelten. Begriffe hierbei sind “mountain”, “country” oder “darling” für Country und Terme wie “fire”, “rebell”, “alone” für Rock.

Insgesamt zeigt sich, dass die menschliche Evaluation wichtige Einblicke und Kontextualisierungen bietet, die für eine umfassende Interpretation von Topic-Modellen unerlässlich sind. Die Schwierigkeit der Aufgabe, Genres anhand der Lyrics von Liedtexten zu erkennen, ist jedoch nicht zu unterschätzen und es fällt uns sehr schwer, bei der Zuordnung überzeugte Entscheidungen zu treffen. Ausnahme hierbei bleibt das Rap Genre, das mit Abstand am meisten einzigartige Lyrics hat und somit fast immer zuordenbar war.

Topic Modell ohne Verben [L]

Die Evaluation des Topic Modells mit Verben zeigt bei der Betrachtung der Topics des gesamten Modells eine deutlich geringere Streuung der Themen mit nur wenigen Ausreißern (siehe Fig. 5).

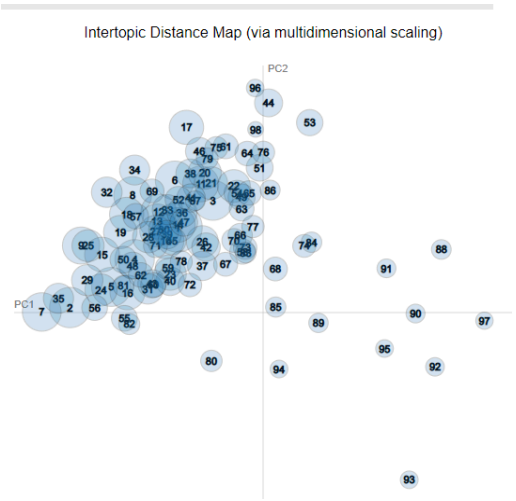


Fig. 5 Verteilung der Topics ohne Verben

Die Themen und dazu entsprechenden Terme sind somit noch ähnlicher und sollten vermuten lassen, dass diese noch schwerer zuzuordnen sind. Dies ist allerdings nicht der Fall und die Zuweisungen der Genres bleiben ähnlich schwer für den Menschen. Es gibt ohne Verben mehr richtungsweisende Wörter, die an Wichtigkeit gewinnen. Ein Beispiel dafür sind Wörter wie “river” oder "honey", die nun beim Identifizieren des Country Genre helfen. Ein Nachteil des Modells ohne Verben ist jedoch, dass ohne Verben wie "love", "sleep" oder "kill" auch viele Informationen verloren gehen, die bei der Analyse genutzt werden könnten.

Bei der Verwendung eines Topic-Modells ohne Verben bleiben die Themen pro Genre im Wesentlichen gleich. Das Rap-Genre lässt sich immer noch am einfachsten identifizieren, aufgrund seiner einzigartigen sprachlichen Merkmale und des vielfältigen Themenbereichs, während die anderen Genres ähnlich schwer zuzuordnen sind.

Die Analyse von ganzen Liedtexten und des damit verbundenen Kontexts ist wesentlich einfacher als die Klassifizierung einzelner Topics, die aus den Lyrics gewonnen wurden. Selbst mit einem verbesserten Verständnis der Terme ohne Verben ist es oft schwierig, eine eindeutige Zuordnung vorzunehmen, da viele Themen und Begriffe in verschiedenen Genres vorkommen können.

Evaluation anhand der Klassifikation [B] [L]

Implementierung und Modell Selektion [B]

Zuerst werden die Trainings- und Testdaten mit Hilfe der Topic Modelle in Verteilungen über die Topics für jede Lyric umgewandelt. Die Trainingsdaten werden anschließend in Trainings- und Validierungsdaten für die Auswahl des Klassifikators getrennt. Als Klassifikatoren werden zum einen ein Regularized Linear Model mit Modified Huber Loss für unterschiedliche Werte von α für die Regularisierung getestet, da dies gut in (Kelechava) funktioniert hat. Außerdem wird eine Logistische Regression getestet. Sowohl für die Daten mit Verben als auch die Daten Ohne war die Logistische Regression die beste Wahl, gemessen anhand Accuracy als auch F1-Score. Das Modell wird schließlich auf den ursprünglichen Trainingsdaten (vor dem Validierungs-Split) trainiert und die Performance auf den Testdaten gemessen. Als Baselines werden verschiedene Text-Klassifikationsmodelle implementiert, darunter Naive Bayes, Logistische Regression und eine Support Vector Machine. Auch hier lieferte Logistische Regression die besten Ergebnisse.

Topic Modell mit Verben [L]

Die Anwendung des Regularized Linear Model mit Modified Huber Loss auf das Topic Modell mit Verben und 100 Topics wurde für verschiedene α getestet und hat den besten F1-Score (Maß für die Genauigkeit eines binären Klassifikators bzw. harmonisches Mittel aus Precision und Recall) mit F1=0.26 bei α = 0.01 gehabt (siehe Fig. 6).

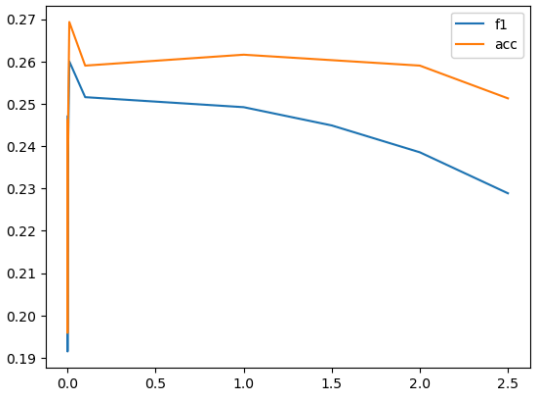


Fig. 6 Graph des F1-Scores in Abhängigkeit von α mit 100 Topics

Dieser Wert ist jedoch sehr gering, wenn man bedenkt, dass der F1-Score bei bloßen Raten sich bei 0.20 einpendeln würde. Ein möglicher Grund könnte die zu feingranulare Aufteilung der Topics sein. Aus diesem Grund wurde der Klassifikator noch mit 50 der Topics getestet. Das Ergebnis zeigte jedoch, dass weniger Topics zu einem noch schlechteren Ergebnis führen und der F1-Score auf lediglich 0.22 kommt (siehe Fig. 7). Das Regularized Linear Model bietet sich somit insgesamt nicht an, um überzeugte Voraussagen zu treffen und findet nur wenig Nutzen in der Topic-Genre Klassifizierung.

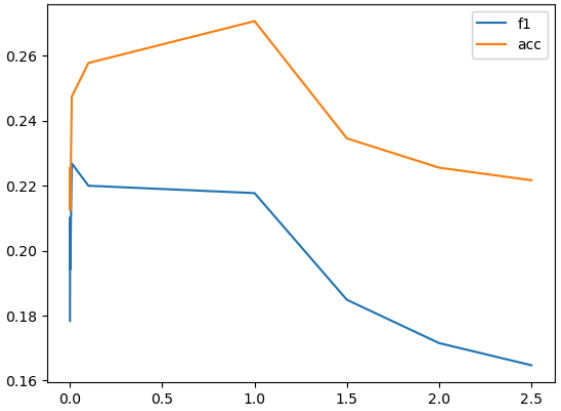


Fig. 7 Graph des F1-Scores in Abhängigkeit von α mit 50 Topics

Eine weitere Herangehensweise ist die Nutzung der logistischen Regression. Diese erzielte einen F1-Score von 0.274 auf den Trainingsdaten und interessanterweise sogar einen F1-Score von 0.372 auf den Testdaten. Der Wert ist immer noch sehr gering und es lässt sich vermuten, dass das Modell zu overfitted war und mit mehr Daten noch besser performen könnte. Wenn man sich nun final die Ergebnisse des Klassifikators pro Genre anschaut, kommt man zu ähnlichen Resultaten wie bei der menschlichen Evaluation.

```
Genre:  pop
percentage of data:  0.18231046931407943
acc score:  0.2524752475247525
Genre:  r-b
percentage of data:  0.21660649819494585
acc score:  0.35
Genre:  country
percentage of data:  0.2157039711191336
acc score:  0.3514644351464435
Genre:  rock
percentage of data:  0.20577617328519857
acc score:  0.2324561403508772
Genre:  rap
percentage of data:  0.1796028880866426
acc score:  0.6683417085427136
```

Fig. 8 Ergebnisse des log. Klassifikators pro Genre mit Verben

Auch der Klassifikator kann Rap-Texte am einfachsten zuordnen und schafft mit einer Accuracy von zwei Dritteln ein gutes Ergebnis. R&B und Country sind deutlich schwerer zu erkennen, jedoch setzen sie sich noch vom Pop und Rock ab, welche nur in geringfügigem Maße das Raten überbieten.

Topic Modell ohne Verben [L]

Bei der Anwendung des Regularized Linear Model mit Modified Huber Loss auf das Topic Modell ohne Verben und 100 Topics gibt es wieder sehr ähnliche Ergebnisse und der F1-Score bewegt sich auch bei 0.25 (siehe Fig. 9).

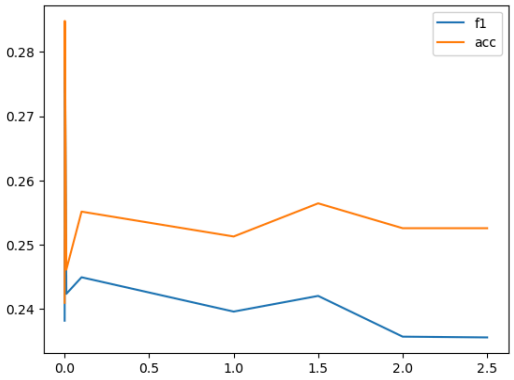


Fig. 9 Graph des F1-Scores in Abhängigkeit von α mit 100 Topics

Fast identisches Verhalten wie beim Klassifikator mit Verben ist bei der Nutzung der logistischen Regression zu beobachten, denn hier liegt der F1-Score ebenfalls bei rund 0.35. Interessante Unterschiede sind bei der Betrachtung der Performance von einzelnen Genres zu finden (siehe Fig. 10).

```
Genre:  pop
percentage of data:  0.18231046931407943
acc score:  0.25742574257425743
Genre:  r-b
percentage of data:  0.21660649819494585
acc score:  0.2916666666666667
Genre:  country
percentage of data:  0.2157039711191336
acc score:  0.38493723849372385
Genre:  rock
percentage of data:  0.20577617328519857
acc score:  0.21052631578947367
Genre:  rap
percentage of data:  0.1796028880866426
acc score:  0.6231155778894473
```

Fig. 10 Ergebnisse des log. Klassifikators pro Genre ohne Verben

Hier ist eine Verbesserung bei der Erkennung von Country Texten zu erkennen mit einem verbesserten Accuracy Score auf 0.384, aber auch eine Verschlechterung beim Klassifizieren von R&B Texten auf einen Accuracy Score von 0.292. Unverändert bleiben Rap mit einer deutlich besseren Genauigkeit und Pop, sowie Rock mit einer Genauigkeit, die dem Raten ähneln.

Weitere Modelle auf den ursprünglichen Daten [L]

Das Modell unter Nutzung des Naive Bayes liefert einen F1-Score von 0.37 und ist somit schon auf gleichem Level mit dem zuvor verwendeten log. Klassifikator.

f1-score 0.3711959003893885

	precision	recall	f1-score	support
pop	0.00	0.00	0.00	137
r-b	0.42	0.69	0.52	173
country	0.36	0.80	0.50	165
rock	0.45	0.08	0.14	174
rap	0.81	0.61	0.70	127
accuracy			0.44	776
macro avg	0.41	0.44	0.37	776
weighted avg	0.40	0.44	0.37	776

Fig. 11 Ergebnisse des Naive Bayes

Rap schneidet auch hier wieder außerordentlich gut ab und zeigt durch hoher Precision und hohem Recall, dass auch außerhalb von Topic Modellen Rap am leichtesten zu Klassifizieren ist. Interessanterweise ist Pop mit einer Precision und einem Recall von 0.00 anscheinend unmöglich zuzuordnen. Obwohl Rock eine hohe Präzision aufweist, wird diese durch den geringen Recall von nur 0.08 beeinträchtigt, was zu einer insgesamt schlechteren Performance führt.

Ein besserer Klassifikator ist der Stochastic Gradient Descent (SGD) Classifier, der einen Gesamt-F1-Score von 0.46 erreicht und somit besser als der Naive Bayes performt.

f1-score 0.46465747238466093

	precision	recall	f1-score	support
pop	0.31	0.15	0.20	137
r-b	0.49	0.53	0.51	173
country	0.45	0.65	0.53	165
rock	0.40	0.30	0.35	174
rap	0.68	0.81	0.74	127
accuracy			0.48	776
macro avg	0.47	0.49	0.46	776
weighted avg	0.46	0.48	0.46	776

Fig. 11 Ergebnisse des SGD Classifier

Das Pop Genre lässt sich hier klassifizieren und auch alle anderen Genres lassen sich ähnlich gut oder besser zuordnen.

Der beste Klassifikator bleibt jedoch der der logistischen Regression mit einem Gesamt-F1-Score von 0.47 und ist somit auch insgesamt der beste Klassifikator für diese Aufgabe.

f1-score 0.47341565869304086

	precision	recall	f1-score	support
pop	0.29	0.23	0.25	137
r-b	0.47	0.51	0.49	173
country	0.54	0.59	0.56	165
rock	0.41	0.29	0.34	174
rap	0.62	0.87	0.73	127
accuracy			0.49	776
macro avg	0.47	0.50	0.47	776
weighted avg	0.46	0.49	0.47	776

Fig. 12 Ergebnisse des Klassifikators mit logistischer Regression auf den Trainingsdaten

Das Verhalten hier ist fast analog zum SGD Classifier und die verbesserte Klassifizierung des Pop Genres qualifiziert die logistische Regression dazu noch, um auf die Testdaten angewendet zu werden (siehe Fig. 13).

f1-score 0.4562994368754324

	precision	recall	f1-score	support
pop	0.25	0.23	0.24	202
r-b	0.46	0.44	0.45	240
country	0.50	0.58	0.54	239
rock	0.37	0.30	0.33	228
rap	0.66	0.79	0.72	199
accuracy			0.47	1108
macro avg	0.45	0.47	0.46	1108
weighted avg	0.45	0.47	0.46	1108

Fig. 13 Ergebnisse des Klassifikators mit logistischer Regression auf den Testdaten

Auch hier bleibt der F1-Score bei soliden 0.46 und zeigt, dass ein Topic Modell mit einem F1-Score von 0.35 deutlich schlechter abschneidet und somit nicht das optimale Modell zur Klassifizierung von Lyrics ist.

Diskussion [L]

Die Ergebnisse dieser Arbeit werfen ein interessantes Licht auf die Komplexität der Genreidentifikation anhand von Themen in Liedtexten und die unterschiedlichen Herangehensweisen von menschlicher Bewertung und automatisierten Klassifikatoren.

Die menschliche Evaluation, als auch die automatisierte Evaluation mit Klassifikator, der Topic-Modelle mit und ohne Verben zeigt, dass bestimmte Themen und Begriffe charakteristisch für verschiedene Musikgenres sind. Das Rap-Genre ist in Hinsicht auf das Ziel der Genreidentifikation an einsamer Spitze. Themen wie Liebe, Drogen, Erfolg und spirituelle Bezüge sind dabei häufig anzutreffen und neben der direkten Sprache ausschlaggebend für Mensch und Klassifikator. Durch häufig verwendete Beleidigungen, Slangwörter oder Begriffe der Black-Community hebt sich das Rap-Genre von allen anderen

ab und kann so auch die besten Erfolge bei der Zuordnung verzeichnen. Diese Begriffe müssen auch die Klassifikatoren erkannt haben, denn das Topic Modell mit log. Regression, als auch die Klassifikatoren wie Naive Bayes oder log. Regression ohne Topic Modell konnten immer einen F1-Score von über 0.6 im Rap-Genre erreichen.
Die beiden Genre R&B sowie Country konnten nur bedingt gut in den Topics erkannt werden. Enthielten diese Themen typische Begriffe über das Landleben oder zum Beispiel alte Kosenamen, so konnten sie leicht dem Country-Genre zugewiesen werden. Bei den meisten Topics war es allerdings der Fall, dass Themen zu viele Überschneidungen hatten und eine Klassifikation entsprechend schwerer war. Das Regularized Linear Model mit Modified Huber Loss, als auch die logistische Regression spiegelt dies wieder mit einem F1-Score von maximal 0.38 bei R&B und Country.
Das Schlusslicht bei der Zuordnung von Themengebieten in Lyrics zu Genres bildeten Pop und Rock. Aufgrund von sehr allgemeinen Themen, wie Liebe und Beziehungen, die zusätzlich wenig komplex sind, war das Pop-Genre von Mensch, als auch Klassifikator schwer zuzuordnen. Im Bereich des Rocks sieht es dabei ähnlich aus, Themen wie Veränderung, Rebellion und Freiheit kommen natürlich auch in anderen Genres vor und ohne eine einzigartige Sprache stechen diese nicht genug hervor, um überzeugte Voraussagen über das Genre zu treffen. Die Klassifikatoren bestätigten dieses Verhalten ebenfalls und so war der F1-Score mit 0.26 immer nur knapp über dem F1-Score 0.20, den man erhalten würde, wenn man immer ein zufälliges Genre wählen würde.

Die Ergebnisse der Klassifikatoren bestätigen größtenteils die Erkenntnisse der menschlichen Evaluation. Der logistische Klassifikator erzielt insgesamt die besten Ergebnisse und kann Rap-Texte am zuverlässigsten zuordnen. Die Genauigkeit der Zuordnung für andere Genres, insbesondere für Pop und Rock, ist jedoch deutlich geringer. Dies legt nahe, dass die Merkmale, die für die Identifizierung von Rap relevant sind, möglicherweise besser von den Klassifikatoren erfasst werden können als die Merkmale anderer Genres. Allerdings ist der Klassifikator nicht in der Lage, subtile Nuancen und Kontexte zwischen den Termen genau zu erfassen, was zu einer Einschränkung bei der Genreerkennung führt.

Ein direkter Vergleich zwischen den Ergebnissen der menschlichen Evaluation und des Klassifikators verdeutlicht die Herausforderungen bei der automatisierten Genreidentifikation, aber auch bei der menschlichen Klassifikation. Während die menschliche Evaluation es ermöglicht, subtile Unterschiede und Kontexte in den Liedtexten zu erfassen, stößt der Klassifikator an seine Grenzen, wenn es darum geht, diese Feinheiten zu erfassen. Grundlage hierfür ist jedoch ein tieferes Verständnis der englischen Sprache und im optimalen Fall auch das Wissen über kulturelle und zeitliche Kontexte. Aus diesem Grund bietet die menschliche Evaluation weiteres Optimierungspotenzial. Durch eine umfassendere Analyse, die über die begrenzten Möglichkeiten dieser Arbeit mit drei Informatikern hinausgeht, könnte ein verbessertes Ergebnis erzielt werden.

Die Analyse zeigt, dass typische Themen und Begriffe zwar Hinweise auf das Genre geben können, aber allein nicht ausreichen, um Genres eindeutig zu identifizieren. Die Vielschichtigkeit der Musiktexte und die unterschiedlichen Interpretationen sowohl von Menschen als auch von Klassifikatoren zeigen, dass die Genreerkennung eine komplexe Aufgabe ist, die weitere Forschung erfordert.

Die Limitationen dieses Ansatzes liegen in der Tatsache, dass Topic-Modelle allein nicht alle Aspekte der musikalischen Identität erfassen können. Andere Merkmale wie Musikstil, Instrumentierung und Gesangstechnik können ebenfalls zur Genreidentifikation beitragen und somit fehlende Information, die oft entscheidend für die Genreidentifikation sind.

Ein möglicher Ansatz zur Verbesserung der Genreerkennung mit Topic Modellen könnte die Integration von zusätzlichen Datenquellen sein, wie z.B. weitere Sprachen neben Englisch. Darüber hinaus könnte die Entwicklung von hybriden Modellen, die menschliche Expertise und automatisierte Techniken kombinieren, zu genaueren Genreerkennungssystemen führen. Diese könnten eine breitere Palette von Musikgenres genau identifizieren und so vielleicht sogar Musikdiensten und Forschern wertvolle Einblicke bieten.

Quellen

Bhattacharjee, Arup, et al., editors. *Machine Learning, Image Processing, Network Security and Data Sciences: Second International Conference, MIND 2020, Silchar, India, July 30 - 31, 2020, Proceedings, Part II*. Springer Nature Singapore, 2020. Accessed 29 March 2024.

Blei, M. D. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993-1022.

Chauhan, S., and P. Chauhan. “Music mood classification based on lyrical analysis of Hindi songs using Latent Dirichlet Allocation.” *2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds*, 2016, pp. 72-76.

Defferrard, Michaël, et al. “FMA: A dataset for music analysis.” *arXiv preprint arXiv:1612.01840*, 2016.

Fell, Michael, and Caroline Sportleder. “Lyrics-based Analysis and Classification of Music.” *ACL Anthology*, 2014, https://aclanthology.org/C14-1059.pdf. Accessed 29 March 2024.

IEEE Staff, and International Conference on Information Technology - the Next Generation IT Summit. *2016 International Conference on Information Technology (InCITe) - The*

Next Generation IT Summit on the Theme - Internet of Things: Connect Your Worlds. IEEE. Accessed 29 March 2024.

Johnson-Roberson, Christopher, and Matthew Johnson-Roberson. “Temporal and Regional Variation in Rap Lyrics.” https://mimno.infosci.cornell.edu/nips2013ws/nips2013tm_submission_25.pdf. Accessed 29 March 2024.

Kelechava, Marc. “Using LDA Topic Models as a Classification Model Input.” *Towards Data Science*, https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28. Accessed 11 March 2024.

Kelechava, Marc. “Using LDA Topic Models as a Classification Model Input.” *Towards Data Science*, https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28. Accessed 29 March 2024.

Laoh, Enrico. “Indonesians' Song Lyrics Topic Modelling Using Latent Dirichlet Allocation.” *IEEE*, 2018, https://ieeexplore.ieee.org/abstract/document/8612562?casa_token=taNPXVIQvbUAAAAA:0Xdk0LGZ0w_pTqyIxXWVGKauwWiiDp5310X6vZGHgk7HBaoHtIhTLDI0b3kC8XPGa9t8MmPxQ7Z7.

Lukic, Alen. “A Comparison of Topic Modeling Approaches for a Comprehensive Corpus of Song Lyrics 1 Introduction.” *Alen Lukic*, http://alenlukic.com/assets/docs/lyric_topic_modeling.pdf. Accessed 29 March 2024.

Miller, John W. “lyricsgenius · PyPI.” *PyPI*, 2021, https://pypi.org/project/lyricsgenius/. Accessed 23 March 2024.

ML Genius Holdings, LLC. “Genius API.” *Genius | Song Lyrics & Knowledge*, https://api.genius.com. Accessed 15 March 2024.

ML Genius Holdings, LLC. “Tags - Music (Genres/Countries/Languages) | Genius.” *Genius*, 2 May 2018,

https://genius.com/Genius-tags-music-genres-countries-languages-annotated. Accessed 17 March 2024.

N. Rubin, Timothy, and et al. “Statistical topic models for multi-label document classification.” *Springer*, 2011, https://doi.org/10.1007/s10994-011-5272-5. Accessed 29 March 2024.

Pedregosa, F., et al. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830.

Radim, R., and Petr Sojka. “Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.” 2010, pp. 45-50.

Rosebaugh, Caleb, and Lior Shamir. “Data Science Approach to Compare the Lyrics of Popular Music Artists.” *Unisia*, vol. 40, 2022, pp. 1-26, https://doi.org/10.20885/unisia.vol40.iss1.art1.

Sasaki, Shoto, and et al. “LYRICSRADAR: A LYRICS RETRIEVAL SYSTEM BASED ON LATENT TOPICS OF LYRICS.” *ISMIR*, 2014, http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/papers/ismir-2014-sasaki.pdf. Accessed 29 March 2024.

Sterckx, Lucas, et al. “Assessing Quality of Unsupervised Topics in Song Lyrics.” *Springer*, 2014, https://link.springer.com/chapter/10.1007/978-3-319-06028-6_55. Accessed 29 March 2024.

Teh, Y., et al. “Hierarchical Dirichlet Processes.” *Journal of the American Statistical Association*, vol. 101, no. 476, 2006, pp. 1566-1581.