Untersuchung des Zusammenhangs zwischen Songlyrics und Genre mit KI

**STUDIENARBEIT**

des Studienganges Informatik

an der Dualen Hochschule Baden-Württemberg Stuttgart

von

Anja Niedermaier

Samara Dominik

Abgabedatum

Bearbeitungszeitraum 01.10.2023 bis 13.06.2024

Matrikelnummer, Kurs XY und 1047506, INF21C

Betreuerin Prof. Dr. Monika Kochanowski

Inhaltsverzeichnis:

1.Einleitung

1.1.Motivation (0,5 Seiten) (Anja)

1.2.Problemstellung/ Forschungshypothesen (1,5 Seiten) (Anja)

1.3.Herangehensweise (1 Seite) (Samara)

2.Theoretische Grundlagen

2.1.Grundlagen der Textverarbeitung und -analyse (2 Seiten) (Samara)

2.2.Techniken der Textanalyse (2 Seiten) (Samara)

2.3.Latent Dirichlet allocation (2 Seiten) (Anja)

2.4.Sentiment-Analyse (2 Seiten) (Anja)

2.5. Wortfrequenzanalyse (mit Zeichensetzung, Liedlänge etc.) (2 Seiten) (Samara)

2.5.Verwendete Tools (1 Seite) (Anja)

2.6.Methoden der Textrepräsentation (3 Seite) (Anja)

2.6.1.Bag of words

2.6.2.Zeilenbasiert

2.6.3.Sliding Windows

2.7.Verwandte Arbeiten (1 Seite) (Samara)

3.Durchführung

3.1.Datenbeschaffung (mit vergleich der anderen Quellen, 1 Seite) (Samara)

3.2.Data Understanding (5 Seiten) (Samara)

3.3.Data Preparation/ Data Cleaning (4 Seiten) (Samara)

3.4.Textanalyse

3.4.1. Latent Dirichlet allocation (5 Seiten) (Anja)

3.4.2. Sentiment-Analyse (5 Seiten) (Anja)

3.4.3. Komination Sentiment und LDA (5 Seiten) (Anja)

3.4.4. Wortfrequenzanalyse (5 Seiten) (Samara)

3.4.5. Analyse der Liedlänge, Wortfülle und Zeichensetzung (5 Seiten) (Samara)

3.4.6.Vergleich der Textanalysemethoden (4 Seiten) (Anja)

3.5 Skalierung der Ergebnisse

3.6.Auswertung Ergebnisse (3 Seiten) (Samara)

4.Fazit (1 Seite) (Anja)

5.Ausblick (1 Seite) (Anja)

# Einleitung

# Theoretische Grundlagen

# Durchführung

## Datenbeschaffung

## Data Understanding

Auch Problem: Reduzierte Anzahl an unterschiedlicher Genres im Vergleich zu der tatsächlichen Menge von Genres

## Data Preparation

tbd

## Data Cleaning

Die Erkenntnisse aus Kapitel 3.2 Data Understanding sollen nun angewendet werden, um den Datensatz bereinigt für die folgenden Analysen zur Verfügung stellen zu können.

Dazu gehören zum einen kleinere Verbesserungen wie das Bereinigen der Duplikate und das Entfernen der nicht-englischen Lieder, sowie von Liedern mit weniger als 50 Zeichen und mehr als 9000 Zeichen, da es sich bei diesen hauptsächlich um fehlerhafte Daten handelt.

Die größte Fehlerquelle, die aus dem Data Understanding ersichtlich wurde, stellen jedoch die teilweise falsch gelabelten Genres dar. Hier muss nun erwogen werden, wie mit diesem Problem in der Datenqualität weiter zu verfahren ist.

Dabei stehen hier einige Möglichkeiten zur Auswahl, zum einen könnten die falsch gelabelten Daten im Datensatz gelassen werden, zum anderen könnten die falsch gelabelten Daten vollständig aus dem Datensatz entfernt, beziehungsweise richtig gelabelt werden. Zwischen diesen beiden Extremen gibt es zusätzliche Optionen, wie die Künstler mit den meisten Liedern für das jeweilige Genre aus dem Datensatz zu nehmen und bei diesen die falsch zugeordneten Genres zu überarbeiten. Zudem könnte auch eine kleinere Mischung aus Liedern von häufigen und weniger häufigen Künstler für jedes der Genre genommen und per Hand richtig gelabelt werden.

Nicht verändert werden soll das Ungleichgewicht der Künstler und Genres.

Im Folgenden soll nun untersucht werden, welche Art der Bereinigung der falsch gelabelten Lieder für das weitere Vorgehen gewählt werden soll. Im Anschluss wird ein Überblick über den bereinigten Datensatz, mit welchem im Folgenden weitergearbeitet wird, gegeben.

### Keine Änderung der falsch zugeordneten Genre

Aus dem zuvor durchgeführten Data Understanding wird ersichtlich, dass einige Künstler in dem Datensatz falsch zugeordnete Genre zu ihren Liedern haben. Diese Erkenntnis stammt allerdings lediglich aus einigen Stichproben, die im Rahmen des Data Understandings durchgeführt wurden. Als Beispiele sind hier, wie bereits im vorherigen Kapitel erwähnt, 50 cent zu nennen, der anscheinend hauptsächlich Pop statt HipHop-Songs im Datensatz vertritt, ähnlich wie auch die Punk-Band Bad Religion hauptsächlich der Pop-Musik und lediglich zu geringen Teilen der Rock-Musik zugeordnet wird.

Wenn man statt diesen Stichproben allerdings die häufigsten Künstler für jedes Genre vergleicht und anschließend recherchiert, welchem Genre diese Künstler tatsächlich hauptsächlich zugeordnet sind, dann kommt man mehrheitlich zu dem Ergebnis, dass die jeweiligen Künstler tatsächlich dem gelabeltem Genre zugeordnet werden können.

Dieser vermeintliche Gegensatz hat zwei Gründe. Einerseits ist es so, dass viele Künstler und ihre Lieder zugleich mehreren unterschiedlichen Genres zugeordnet werden können. Dieser Datensatz ordnet jedem Lied aber immer nur genau ein Genre zu, sodass verschiedene Genremischungen und Zuordnungsmöglichkeiten nicht berücksichtigt wurden. Andererseits muss beachtet werden, dass Genres wie beispielsweise die Rockmusik einige Unterformen haben können, so kann man die Punk- oder die Alternative-Musik beispielsweise als eine Unterform der Rock-Musik ansehen. Je differenzierter die Labeling eines Liedes, desto genauer werden natürlich die Analyseergebnisse. Allerdings ist es in diesem Sinne auch nicht falsch, Punk-Musik als Rock-Musik zu bezeichnen, da diese schließlich das übergreifende Musikgenre ist. Des Weiteren wandelt sich Musik und die Wahrnehmung dieser im Laufe der Zeit. Während Elvis Presley beispielsweise in den 50ern einer der populärsten Künstlern war, der auch neue, revolutionäre Ideen habe und zu dieser Zeit durchaus dem Rock zugeordnet wurde, so wäre seine Musik heutzutage eher einer Unterform des Rocks, dem Rock’n’Roll oder weiteren Genres wie dem Blues oder dem Country zuzuordnen.

Diese Betrachtung führt einen zu dem Schluss, dass die meisten Lieder, bis auf einige Ausnahmen, im Datensatz nicht falsch, sondern höchstens unvollständig gelabelt sind. Die weitere Analyse also über den gesamten Datensatz durchzuführen hätte die Vorteile, dass mit dem geringsten Arbeitsaufwand an Bereinigung der Daten und dabei dem größten Datensatz weitergearbeitet werden kann. Als Nachteil dieses Vorgehens ist dennoch die Möglichkeit leicht verfälschter oder nicht ganz genauer Ergebnisse zu nennen.

### Vollständiges Entfernen/ Korrigieren der falsch zugeordneten Genre

Mit der vollständigen Entfernung oder Korrektur der Daten kann das Problem der falschen, beziehungsweise weniger genauen Ergebnisse gelöst werden. Hierfür müsste allerdings jeder Datensatz einzeln durchgegangen und für jedes Lied recherchiert werden, ob es sich um das richtig zugeordnete Genre handelt. Bei der Entscheidung für das Entfernen der falsch zugeordneten Lieder müssten im nächsten Schritt alle betroffenen Datensätze gelöscht werden. Wenn sich zu einer Korrektur der betroffenen Datensätze entschieden wird, so sollten weitere Genres dem Datensatz hinzugefügt werden, um eine bessere Differenzierung und genauere Ergebnisse erreichen zu können. Von einer Zuordnung von mehreren Genres zu einem Song sollte auch hier abgesehen werden, da dies Probleme in der späteren Textanalyse verursachen könnte. Der Vorteil des Löschens aller fehlerhaften oder ungenauen Datensätze wäre, dass mit einem fehlerfreien Datensatz gearbeitet werden könnte. Die Nachteile sind jedoch die zeitaufwendige Arbeit des Herausfindens der fehlerhaften Datensätze, zumal diese Aufgabe nur schwer automatisiert werden kann, sowie die Tatsache, dass bei dieser Methode nicht bekannt ist, wie viele Datensätze nach der Bereinigung übrigbleiben. Dagegen kann beim Korrigieren der fehlerhaften Datensätze nach wie vor mit dem vollständigen Datensatz weitergearbeitet werden, zudem würde diese Bereiningungsmethode einen noch genaueren Datensatz und damit möglicherweise bessere Ergebnisse hervorbringen. Allerdings wäre dieses Vorgehen deutlich zeitaufwendiger als das reine Löschen der fehlerhaften Datensätze, welches allein schon sehr viel Arbeitszeit kostet.

### Überarbeitung der falsch zugeordneten Genre der häufigsten Künstler

Das Problem des hohen Arbeitsaufwandes in der Bereinigung aller Datensätze kann dadurch gelöst werden, dass von vorneherein bereits nur ein kleinerer Teil des Originaldatensatzes betrachtet und später auch für die Analysen genutzt wird. Eine Möglichkeit diesen Teil festzustecken, ist die 20 häufigsten Künstler jedes Genres herauszufiltern, was 100 unterschiedliche Künstler im Datensatz und 30- 35 tausend unterschiedliche Lieder bedeuten würde. Anschließend würde jeder Künstler einzeln betrachtet und mit allen seinen Liedern, dem Hauptgenre des Künstlers zugeordnet werden. Dabei können auch neue Genres hinzugefügt werden. Der Vorteil dieser Methode ist, dass innerhalb des verkleinerten Datensatzes jeder Künstler und die zugeordneten Genres einzeln überprüft werden können, ohne einen übermäßigen Arbeitsaufwand zu produzieren. Der Datensatz bleibt dabei dennoch bei einer ausreichend großen Größe. Allerdings kann hier aufgrund des nach wie vor großen Datensatzes nicht jedes Lied einzeln durchgegangen werden, sodass, vor allem bei Künstlern, die ihr Genre im Laufe ihrer Karriere geändert haben, neue Fehler erzeugt, beziehungsweise sogar ursprünglich richtig gelabelte Datensätze falsch gelabelt werden könnten. Zudem ist nicht auszuschließen, dass wir keine vorbeeinflussten Ergebnisse in der Analyse erzielen, wenn wir für den neuen Datensatz lediglich die 20 häufigsten Künstler jedes Genres auswählen.

### Überarbeitung der falsch zugeordneten Genre aus eine Mischung von Künstlern

Das Problem der vorbeeinflussten Werte könnte wiederum durch eine zufällige Auswahl 30 unterschiedlicher Künstler aus jedem Genre gelöst werden. Die Anzahl der Künstler aus jedem Genre wurde um 10 erhöht, da bei einer zufälligen Auswahl nicht garantiert ist, ob ein Künstler lediglich ein oder mehrere hunderte Lieder im Datensatz hat und somit die Mindestanzahl an Liedern erhöht wird. Somit wären bei diesem Vorgehen 300 Künstler einzeln, wie in der zuvor beschriebenen Methode, zu untersuchen und mögliche fehlerhafte Datensätze entsprechend zu beschriften. Dies entspricht im schlechtesten Fall 300 Datensätzen und kann im besten Fall 30 bis 40 tausend Datensätze beinhalten. Der Vorteil dieser Methode deckt sich mit dem der vorherige, zudem ist hier durch die zufällige Auswahl der Künstler die Wahrscheinlichkeit an vorbeeinflussten Ergebnissen deutlich geringer.

### Auswahl der Vorgehensweise

Nachdem in den vorherigen Kapiteln nun verschiedenen Vorgehensweisen bezüglich der falsch gelabelten Genres untersucht wurden, wird sich nun dafür entschieden, den Datensatz hinsichtlich der falsch zugeordneten Genres in seiner Originalversion beizubehalten. Der Grund hierfür ist die Tatsache, dass nach der Analyse deutlich wurde, dass nur eine geringe Anzahl der Genres tatsächlich falsch zugeordnet ist und die restlichen, vermeintlich falsch zugeordneten Lieder lediglich nicht ganz genau zugeordnet wurde. Würden wir nun diese Lieder korrigieren oder entfernen, wäre das zu zeitaufwendig für den Rahmen der Projektarbeit und nur einen Teil des Datensatzes über die Künstler und deren Hauptgenres zu überarbeiten, birgt ein zu hohes Risiko mehr neue Fehler zu erzeugen, als es korrigieren würde.

### Bereinigter Datensatz

Die Bereinigung des Datensatzes beinhaltet nun gemäß der Ergebnisse des Data Understandings das Entfernen der Duplikate, sowie aller länger die kürzer als 50 und länger als 9000 Zeichen sind. Davon abgesehen soll der Datensatz so groß wie möglich bleiben.

Nach der Bereinigung…

## Textanalyse

## Skalierung der Ergebnisse

## Auswertung der Ergebnisse

# Fazit

# Ausblick