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

## 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 könnte dieses Problem taktisch ignoriert werden

Nicht verändert werden soll das Ungleichgewicht der Künstler und Genres.

## Textanalyse

## Skalierung der Ergebnisse

## Auswertung der Ergebnisse

# Fazit

# Ausblick