

Dispositionspapier zur Studienarbeit

Untersuchung des Zusammenhangs zwischen Songlyrics und Genre mit KI

Anja Niedermeier, Samara Dominik

11.01.2024

1 Kurzbeschreibung der Arbeit

Bei der Betrachtung von Musik entstehen häufig unterschiedliche Standpunkte. Eine weitgehend akzeptierte Ansicht ist jedoch, dass bestimmte Musikgenres tendenziell immer wieder dieselben Themen in ihren Songtexten behandeln. Diese weit verbreitete Meinung soll in der Arbeit, mithilfe von künstlicher Intelligenz und maschinellen Lernmethoden näher beleuchtet und analysiert werden.

In diesem Zusammenhang sollen folgende Hypothesen untersucht werden:

- Das Thema eines Liedes lässt sich durch Sprachverarbeitungsmethoden erschließen
- Es besteht ein Zusammenhang zwischen Lyric und Genre
- Anhand der Lyrics eines Liedes lässt sich das Genre mit einer Genauigkeit von 70% erschließen
- Anhand der Lyrics lässt sich schließen welche Genres thematisch ähnlich sind
- In populären Genres (Rock und Pop) gibt es eine größere Diversität an Textthematiken

Um diese Hypothesen belegen oder widerlegen zu können, müssen einige Bedingungen gegeben sein. So wird ein Datensatz an Liedern mit Liedtexten und Genres benötigt, über welchen die Analysen laufen sollen. Um möglichst genaue Aussagen treffen zu können, soll dieser Datensatz so groß wie möglich sein, was in der späteren Analyse die benutzten handelsüblichen Computer an ihre Grenzen bringen könnte, weswegen hier ein leistungsfähiger Computer der Hochschule gestellt werden soll. Zudem werden theoretisches Wissen, sowie methodische Fähigkeiten der Textanalyse und der Sprachverarbeitung benötigt, um die grundsätzlichen Themen eines Liedes aus dem Liedtext für die weitere Verarbeitung erschließen zu können. Hierfür werden in dieser Arbeit die Latent Dirichlet allocation, sowie die Sentiment-Analyse verwendet. Deren Ergebnisse sollen später über verschiedene Techniken wie Clustering oder visuelle Darstellung untersucht werden.

2 Gliederung und Zeitplan

Diese Studienarbeit soll sich an der folgenden Gliederung orientieren:

1. Einleitung
 - 1.1. Motivation
 - 1.2. Problemstellung
 - 1.3. Herangehensweise
2. Theoretische Grundlagen
 - 2.1. Grundlagen der Textverarbeitung und -analyse
 - 2.2. Techniken der Textanalyse
 - 2.3. Latent Dirichlet allocation

- 2.4. Sentiment-Analyse
- 2.5. Verwendete Tools
- 2.6. Methoden der Textrepräsentation
 - 2.6.1. Bag of words
 - 2.6.2. Zeilenbasiert
 - 2.6.3. Sliding Windows
- 2.7. Verwandte Arbeiten
3. Durchführung
 - 3.1. Datenbeschaffung
 - 3.2. Data Understanding
 - 3.3. Data Preparation
 - 3.4. Textanalyse
 - 3.4.1. Anwendung Latent Dirichlet allocation
 - 3.4.2. Sentiment-Analyse
 - 3.4.3. Vergleich und Korrelationen
 - 3.5. Auswertung Ergebnisse
4. Fazit
5. Ausblick

Die Bearbeitung der Arbeit soll sich dabei an der Gliederung orientieren, wobei das Augenmerk primär auf die Ausführung des praktischen Teils gelegt wird und die Bearbeitung der theoretischen Grundlagen parallel und nach der Erarbeitung des praktischen Teils erfolgen soll. In der unteren Grafik des Meilensteinplans werden die benötigten Schritte dem Zeitplan zugeordnet.

Projektbeginn				
Datensammlung	Anja Niedermeier, Samara Dominik	100%	1.10.23	24.10.23
Methodensammlung	Anja Niedermeier, Samara Dominik	100%	1.10.23	17.10.23
Auswertungsmöglichkeiten	Anja Niedermeier, Samara Dominik	100%	1.10.23	31.10.23
Gliederung erstellen	Anja Niedermeier, Samara Dominik	100%	1.10.23	6.11.23
Besprechung mit Betreuerin	Anja Niedermeier, Samara Dominik	100%	6.11.23	6.11.23
Data Understanding				
Datenbetrachtung	Samara Dominik	100%	6.11.23	6.12.23
Data Preprocessing	Samara Dominik	100%	6.11.23	6.12.23
Data Understanding	Samara Dominik	100%	6.11.23	6.12.23
Erstes Methodentesting	Samara Dominik	100%	6.11.23	6.12.23
Dokumentation der Ergebnis	Anja Niedermeier, Samara Dominik	100%	6.11.23	6.12.23
Besprechung mit Betreuerin	Anja Niedermeier, Samara Dominik	100%	6.12.23	6.12.23
Dispositionspapier	Anja Niedermeier, Samara Dominik	100%	6.12.23	12.1.24

Textanalyse				
Data Preparation	Anja Niedermeier, Samara Dominik	0%	12.1.24	1.3.24
Latent Dirichlet allocation	Anja Niedermeier, Samara Dominik	0%	12.1.24	1.3.24
Sentiment-Analyse	Anja Niedermeier, Samara Dominik	0%	12.1.24	1.3.24
Dokumentation der Ergebnis	Anja Niedermeier, Samara Dominik	0%	12.1.24	1.3.24
Besprechung mit Betreuerin	Anja Niedermeier, Samara Dominik	0%	?	?
Auswertung und Überarbeitung				
Auswertungen, Vergleiche	Anja Niedermeier, Samara Dominik	0%	1.3.24	10.3.24
Ggf. Anpassung von Daten	Anja Niedermeier, Samara Dominik	0%	10.3.24	24.3.24
Nötige Überarbeitungen	Anja Niedermeier, Samara Dominik	0%	10.3.24	24.3.24
Dokumentation der Ergebnis	Anja Niedermeier, Samara Dominik	0%	1.3.24	24.3.24
Besprechung mit Betreuerin	Anja Niedermeier, Samara Dominik	0%	?	?
Schriftlicher Teil				
Einleitung	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Grundlagen der Textverarbe	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Latent Dirichlet allocation	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Sentiment-Analyse	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Verwendete Tools	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Datenformate	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Verwandte Arbeiten	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Ausarbeitung der Durchführ	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Fazit und Ausblick	Anja Niedermeier, Samara Dominik	0%	24.3.24	1.6.23
Besprechung mit Betreuerin	Anja Niedermeier, Samara Dominik	0%	?	?
Abgabe der Arbeit	Anja Niedermeier, Samara Dominik		1.10.23	13.6.24

3 Grundlegende Literatur

Durch die Studienarbeit zum Thema «Untersuchung des Zusammenhangs zwischen Songlyric und Genre mit KI» soll die Frage beantwortet werden, ob das allgemeine Gefühl der meisten Menschen stimmt, dass Liedtexte und Genre nicht nur zusammenhängen, sondern Genres tatsächlich jeweils hauptsächlich ein bis zwei Themen, die aus den Liedtexten herausfilterbar sind, behandeln.

Die Idee, den Zusammenhang zwischen Text und Genre zu untersuchen ist dabei nicht neu, so haben de Lima, A.A., Nunes, R.M., Ribeiro, R.P., Silla dieses Thema bereits 2014 in ihrer Arbeit «Nordic Music Genre Classification Using Song Lyrics» (DOI: https://doi.org/10.1007/978-3-319-07983-7_14) auf Basis eines Datensatzes nordischer Musik untersucht. Dabei wurde auch ein Augenmerk auf die verschiedenen Ansätze der Analysemethoden gelegt und zwischen sprachunabhängigen wie den n-grams, sowie

sprachabhängigen Methoden unterschieden. Im Gegensatz zu dieser Arbeit, nutzt unsere Arbeit einen größeren Datensatz, der mehrere Länder abdeckt, außerdem verwenden und vergleichen wir anderen Analysemethoden wie beispielsweise die Latent Dirichlet allocation. Zusätzlich zur eben beschriebenen Arbeit haben auch zwei Studierende, Boonyanit und Dahl, in Stanford die Arbeit «Music Genre Classification using Song Lyrics» geschrieben, in welcher sie verschiedene Techniken der Zuordnung eines Genres zu einem Text basierend auf dessen Lyric untersuchen. Das gleiche Thema wird auch in Tsaptsinos, ebenfalls aus Stanford, in der Arbeit mit dem Titel «Lyrics-Based music genre classification using a hierarchical attention network» und Junru Yang, der University of Victoria, in der Arbeit «Lyric-Based Music Genre Classification» behandelt. Um dies zu tun, muss von vorneherein ausgegangen werden, dass Genre und Liedtext zusammenhängen. In diesem Kontext stellt unsere Arbeit also die Grundlage für die Arbeit der vorangegangenen Arbeiten, indem wir untersuchen, ob diese Annahme grundsätzlich zutrifft.

So lässt sich zusammenfassen, dass die bisherigen Papers sich hauptsächlich darauf konzentrieren direkt Genre zu Lyrics zuzuordnen und dabei versuchen einander mit verschiedenen KI-basierten- und nicht KI-basierten-Verfahren, sowie in der Auswahl der untersuchten Aspekte (siehe auch Mayer, Neumayer, Rauber «Rhyme and Style Features for Musical Genre Classification by Song Lyrics», 2008) in ihrer Genauigkeit zu übertreffen. Dagegen soll in dieser Studienarbeit mithilfe eines größeren Datensatzes über mehrere Herkunftsländer und verschiedenen Textanalyse, sowie Auswertungsmethoden bewiesen werden, dass die Annahme, auf welche die anderen Arbeiten aufsetzen, wahr ist.