HAUSARBEIT FÜR WPM FORSCHUNGSPROJEKT TEXT ANALYTICS

Erstellung eines Entscheidungsbaumes zur Klassifikation von Hits in den deutschen Single-Charts in Jahrzehnten seit 1980

18. Dezember 2020

Tobias Kück (9006) I18d
Joshua Mähl (9300) I18d
Danny Marcel Kröger (9403) I18d

Erstellung eines Entscheidungsbaumes zur Klassifikation von Hits in den deutschen Single-Charts in Jahrzehnten seit 1980

Danny Marcel Kröger, Joshua Mähl und Tobias Kück

Nordakademie, Wirtschaftsinformatik, Elmshorn, Deutschland
{danny_marcel.kroeger, joshua.maehl, tobias.kueck}@nordakademie.de

Abstract. Der vorliegende Artikel gibt einen Überblick über das Hit-Potenzial eines Songs in einem bestimmten Jahrzehnt. Als Analysebasis wurden dazu die offiziellen deutschen Single Charts zwischen 1980 und 2020 herangezogen, wobei der Fokus auf Liedern mit deutschen Texten liegt. Diese wurden hinsichtlich ihrer Top 3 Themen, der Länge in Sekunden, der Anzahl der Wörter sowie der Wortvarianz analysiert. Dadurch lässt sich ein Lied mit gegebenen Features unter Zuhilfenahme eines Entscheidungsbaumes als ein Hit oder Nicht-Hit klassifizieren. Das entstandene Modell ist jedoch nur bedingt aussagekräftig, da die Klassifikation eines Liedes als Nicht-Hit zuverlässiger gelingt als die eines Hits.

Keywords: DASC-PM, Entscheidungsbaum, Hit-Klassifikation

1 Einleitung

Der Musikgeschmack ist ein oft umstrittenes Thema und auch die Diskussion, bei welchen Liedern es sich um einen sogenannten Hit handelt, kann schnell Meinungsverschiedenheiten auslösen. Der durchschnittliche Musikgeschmack des Deutschen ist in den offiziellen deutschen Charts manifestiert, die seit 1977 wöchentlich eine Liste mit den derzeit beliebtesten Liedern veröffentlichen. Darin enthalten sind sowohl deutsch-, als auch fremdsprachige Lieder. Die Frage, welche Merkmale ein Lied zu einem Hit macht, beantwortet sie allerdings trotzdem nicht. Zwar existieren Forschungen auf dem Gebiet, diese fokussieren sich allerdings auf ein internationales Publikum und geben keine Auskunft über deutsche Lieder.

Der vorliegende Artikel soll deshalb die Frage beantworten, wie bestimmte Merkmale eines Songs aussehen müssen, damit dieser als ein Hit klassifiziert werden kann. Die Datenbasis dazu bilden die deutschsprachigen Lieder der deutschen Charts zwischen 1980 bis 2020. Außerdem werden die Daten in Jahrzehnte aufgeteilt. Die Songs in den dadurch entstehenden vier Arbeitsgrundlagen werden anhand der folgenden Merkmale analysiert: Thema des Textes, Gesamtanzahl der Wörter, Anzahl der unterschiedlichen Wörter und Liedlänge in Sekunden. Mithilfe dieser

Charakteristiken wird ein Entscheidungsbaum zur Klassifikation eines Liedes als Hit oder Nicht-Hit erstellt.

Dazu teilt sich die Arbeit in unterschiedliche Kapitel, beginnend mit der Untersuchung verwandter Arbeiten und der Erklärung der verwendeten Methodologie. Anschließend folgt die Datenerhebung und -bereinigung, die Themenanalyse und das Feature Engineering. Aus den so gewonnenen Daten lässt sich im nächsten Abschnitt ein Entscheidungsbaum bilden, welcher einen gegebenen Song mithilfe der zuvor genannten Daten als ein Hit oder Nicht-Hit klassifizieren kann. Schließlich folgt die Betrachtung der Ergebnisse, dessen Interpretation sowie ein abschließendes Fazit.

2 Stand der Forschung

Im Folgenden werden themennahe und verwandte Forschungen vorgestellt, um einerseits den State-of-the-Art zu definieren und andererseits dieses Experiment an die bereits erforschten Erkenntnisse anzupassen. Hierfür wird Literatur verwendet, deren Qualität durch den Peer-Review-Prozess bestätigt wurde.

Forschungen im Bereich Musik sind keine Neuheit, auch nicht für die Disziplin Data Science. Zangerle et al. stellen einen auf Deep-Learning-basierten Ansatz zur Vorhersage von potenziellen Hit-Songs mittels Audio Features vor [1, S. 319]. Sie definieren einen Hit als ein Lied, dass zu einem beliebigen Zeitpunkt in den Billboard Top 100 präsent war [1, S.319]. Eines ihrer Key Findings ist, dass neben den grundlegenden Audio Features auch das Release-Jahr, Sentiments und Thematiken des Stückes einen positiven Effekt auf das Klassifikationsresultat haben [1, S. 319].

Eine weitere Ansatzmöglichkeit zur Analyse von Liedern präsentieren Fell und Sporleder. Ihr Experiment reduziert sich auf verschiedene Klassifikationen und Analysen von Liedtexten [2, S. 620]. In der vorgestellten Feature-Auswahl wird u. A. auch die Länge des Liedes als relevantes Klassifikationsmerkmal gedeutet [2, S. 622]. Insbesondere die Liedthemen und die jeweiligen Längen seien die effizientesten Features für diverse Klassifikationen [2, S. 625-627].

Pachet und Roy forschen auf einer Metaebene über die Interpretation der Validität einer „Hit Song Science“. Sie untersuchen Hit Charakteristiken und kombinieren hierfür akustische und „menschliche“ Features, wie u. A. Liedthema, Liedlänge und Wortvarianz und -anzahl [3, S. 356-357]. Kim et al. klassifizieren Hits auf Basis verwandter Hashtags [4, S. 51]. Eine der Hauptaussagen bezieht sich auf die Subjektivität der Definition eines Hits [5, S. 52].

Einige der untersuchten Forschungen, die sich auf den Bereich Musik beschränken und Themen identifizieren, wählen eine Modellierung mittels Latent Dirichlet Allocation (LDA). LDA wurde 2003 von Blei et al. publiziert und wie folgt definiert: „Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words“ [5, S. 996]. Die Autoren beschreiben folgende mathematische Funktionsweise: Man wähle eine feste Themenanzahl N und bilde eine Dirichlet-Verteilung über alle Terme eines Dokumentes. Dann bilde man N Multinomialverteilungen über die beschriebene

Dirichlet-Verteilung und erhalte pro Thema z_N mindestens ein Wort w mit der höchsten bedingten Wahrscheinlichkeit $P(w|z_N)$ [5, S. 996]. Dieser Prozess wird wiederholt, wobei sich die Wahrscheinlichkeiten nach jeder Iteration anpassen.

Die Effizienz dieses Wahrscheinlichkeitsmodells für ein automatisiertes Topic Modelling wird von Teneva und Cheng im Kontext der Keyphrase Extraction untersucht. Die Forscher identifizieren einige Vorschläge für potenzielle Effizienzsteigerungen, wie z. B. ein Dokument nur einmal zu untersuchen [6, S. 530]. Dennoch werde LDA als geltende State-of-the-Art-Lösung im Topic Modelling wahrgenommen [6, S. 534]. Eine ähnliche Untersuchung führen Suzuki und Fukumoto durch, die LDA zur Topic Detection von Dokumentzusammenfassungen verwenden und als effektive Methode beschreiben [7, S. 241].

Ein wichtiger Aspekt von Text Mining Projekten ist das Preprocessing der zu analysierenden Daten. Rianto et al. untersuchen die Accuracy einer Text-Klassifikation unter Zuhilfenahme verschiedener Stemming Methoden [8, S. 1]. Eine inadäquate Stemming-Methode habe einen negativen Effekt auf die Accuracy [8, S. 12]. Boudin et al. führen eine ähnliche Untersuchung der Performance verschiedener Preprocessing Techniken durch. Sie heben hervor, dass gewöhnliche Preprocessing Methoden (z. B. Stoppworte entfernen) insbesondere für das Extrahieren von Themen empfohlen wird [9, S. 126].

In verwandten Forschungen werden des Öfteren Entscheidungsbäume zur Klassifikation herangezogen. Grąbczewski schreibt, dass der meistverwendete Entscheidungsbaum-Ansatz eine Induktion per rekursiven Top-Down-Splits vorsieht und jeder Split auf dem Prinzip „Purity Gain“ basiert [10, S. 12]. Der „Purity Gain“ entspricht mathematisch bekannten Reinheitsmaßen. So basiert der ID3-Algorithmus auf der Entropie [10, S. 14] und der CART-Algorithmus auf dem Gini-Index [10, S. 16] als Reinheitsmaß. CART sei sowohl für stetige und diskrete Feature geeignet [10, S. 16]. Boros et al. weisen auf den positiven Lerneffekt für die Daten durch das Verwenden von Entscheidungsbäumen hin [11, S. 103].

3 Methodologie

Es existieren einige Prozessmodelle für datenanalytische Projekte. Ein übergreifend bekanntes und universell eingesetztes Modell ist CRISP-DM, welches im Jahr 2000 durch ein von der Europäischen Union gefördertes Konsortium publiziert wurde [12, S. 3]. Eine von den Herausgebern beschriebene Besonderheit des Prozesses ist die Unabhängigkeit von der jeweiligen Domäne [12, S. 3]. Hingegen gibt es Kritiker, die dieses Modell, z. B. aufgrund des Alters, ablehnen.

Schulz et al. greifen mehrere Kritikpunkte an existierenden Data Mining Vorgehensweisen auf, darunter auch CRISP-DM. Die Autoren weisen auf eine Einschränkung des CRISP-DM auf den Unternehmenskontext hin [13, S. 14]. Außerdem biete das Modell keine Verknüpfungen mit nahen Elementen wie Arbeitsweisen und -werkzeugen sowie Teamrollen [13, S. 13]. Sie identifizieren eine Notwendigkeit, die „Datenanalyse um ein geeignetes Vorgehensmodell zu ergänzen“ [13, S. 8].

Schulz et al. wandeln die beschriebene Notwendigkeit in ein Modell um, das die Autoren Data-Science-Process-Model (DASC-PM) nennen. Im Folgenden wird der Prozess abgebildet:

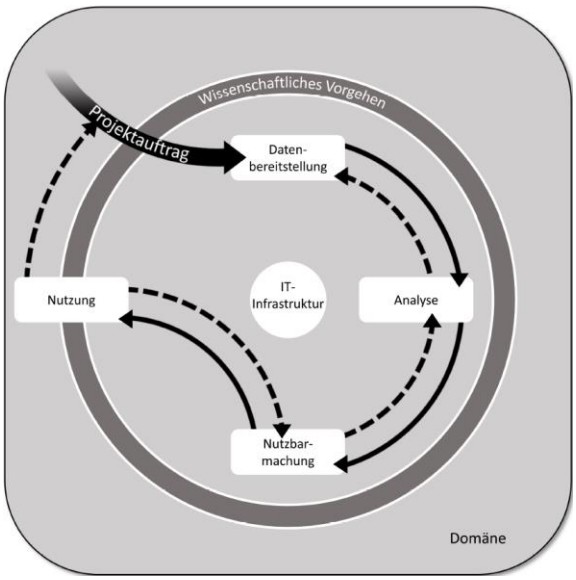

Abbildung 1. Data-Science-Vorgehensmodell DASC-PM [13, S. 23]

Das abgebildete Modell bietet eine für diese Untersuchung optimale Prozessgrundlage, da die Einbettung in die untersuchte Domäne (Deutsche Single Charts) zu jedem Zeitpunkt gegeben sein muss. Ebenfalls ist ein wissenschaftliches Vorgehen aufgrund der verbundenen Transparenz wichtig. Durch die vier Schritte Datenbereitstellung, Analyse, Nutzbarmachung und Nutzung werden sämtliche CRISP-DM Schritte aufgefangen und erweitert, indem Rückgänge bei notwendigen Anpassungen zu vorherigen Schritten möglich sind (siehe Abbildung 1).

Die untersuchte Literatur bietet eine valide Grundlage, um Features zu definieren, die einen Hit in einem gegebenen Jahrzehnt charakterisieren können. Wie mehrere Forschungen nahelegen, ist zunächst eine Topic Detection vorgesehen (siehe Kapitel 2). Andererseits konnte auch die Liedlänge als erfolgsversprechendes Feature identifiziert werden. Zudem wird die Wortvarianz als mögliches Merkmal zitiert, weshalb dies neben der absoluten Wortmenge in einem Lied als weiteres Feature in diesem Experiment betrachtet wird. Da ein Überfluss an eventuell inkompatiblen Features zu einem Effizienzabfall des Modells führen könnte und der gegebene Rahmen limitiert ist, werden weitere mögliche zitierte Liedmerkmale wie Melodien oder Popularitätsgrade des/der Interpreten nicht in die Analyse aufgenommen.

Vor dem eigentlichen Experiment erfolgt laut DASC-PM der Projektauftrag (siehe Abbildung 1). In diesem Kontext konnte das Forschungsprojekt frei gewählt werden. Zur Beantwortung der vorgestellten Fragestellung bedarf es zusätzlich einer Definition des Begriffes „Hit“. Einige Autoren erkennen ein Lied, das jemals in den Top 100

Charts präsent war, als Hit an (siehe Kapitel 2). Andererseits wird von Kim et al. eine individuelle Definition nahegelegt [4, S. 52], falls das Experiment dies benötigt. In diesem Paper wird folgende Definition verwendet: Ein Hit ist ein Lied, das sich mindestens zwei Wochen in den Top 20 der Deutschen Single Charts etablieren konnte.

Der DASC-PM Schritt Datenbereitstellung beinhaltet das Anfertigen von SQL-Abfragen zur Exportierung der Daten aus der Nordakademie-Datenbank, wo sämtliche Liedinformationen gespeichert sind. Da keine Informationen zu der Liedlänge in der genannten Datenbank existieren, werden die Daten durch ein individuelles Programm eingelesen, welches ein gegebenes Lied in einem Browser sucht und die Längeninformation semi-automatisch extrahiert.

Für die Themenanalyse ist insbesondere der Umgang mit Liedtexten notwendig. Nach einem Preprocessing wird wie in den meisten State-of-the-Art-Forschungen (siehe Kapitel 2) LDA als Algorithmus zur Topic Detection eingesetzt. Nach Anwendung der LDA wird aufgrund der großen Menge an resultierenden Themen ein deutschsprachiges, vortrainiertes Word Embedding Modell angewendet, um Ähnlichkeit zu abstrahieren und ein Clustering mittels des k-Means Algorithmus in definierte Themenbereiche zu ermöglichen. k-Means ist ein Algorithmus, bei dem k zufällige Mittelpunkte initialisiert werden, die Zugehörigkeit von Themen zu Clustern aufgrund von relativ „nahen“ Vektoren bestimmt wird und die Mittelpunkte bei jeder Iteration angepasst werden, sodass sich die Distanzen der Vektoren zum jeweiligen Cluster-Mittelpunkt minimieren.

Ein weiterer im DASC-PM Schritt Analyse beinhalteter Aspekt ist die Klassifikation. Um das komplexe Zusammenspiel der Komponenten eines Liedes darstellen zu können, bietet sich eine Klassifikation anhand eines Entscheidungsbaumes unter Verwendung des CART-Algorithmus (siehe Kapitel 2) an.

Damit die genannten Vorbereitungs- und Analyseschritte durchgeführt werden können, ist eine Data-Mining-Analyseplattform notwendig (IT-Infrastruktur in DASC-PM). In diesem Experiment wird das Programm KNIME verwendet. Valide Gründe für die Auswahl dieses Programms sind: Es bietet eine interaktive Benutzeroberfläche, die auf relativ simple Art den Datenfluss visualisiert. Außerdem ist ein Überprüfen des Prozessstatus bei jedem existierenden Knoten möglich, was für mögliche Zyklen von DASC-PM Schritten unabdingbar ist. Das DASC-PM wird durch die Schritte Nutzbarmachung und Nutzung abgeschlossen. Diese Schritte werden durch die folgenden Diskussionsabschnitte in diesem Artikel abgedeckt.

4 Datenbereitstellung

4.1 Datenerhebung

Als Ausgangspunkt für die Datenerhebung wurde die Webseite offiziellecharts.de verwendet. Jedem Lied wird dabei eine eindeutige ID zugewiesen. Die Webseite dokumentiert seit 1977 wöchentlich deren Chart-Position.

Somit ist die Basis für eine weitere Anreicherung der Daten geschaffen. Da es sich bei der Arbeit grundsätzlich um eine Analyse des Liedtextes handelt, mussten auch

diese Daten erhoben werden. Hierfür wurde der Lied-Titel und Interpret in einem Browser gesucht und das erste Suchergebnis für den Liedtext in die Datenbank übernommen. Dabei handelte es sich um unterschiedliche Webseiten, die als Datenbasis dienten. Hier muss sich allerdings bewusstgemacht werden, dass die Daten meistens nicht von den Interpreten selbst veröffentlicht werden und daher einige Unreinheiten enthalten können.

Ein weiteres Merkmal ist die Länge eines Liedes in Sekunden. Dieses musste ebenfalls erhoben werden und soll Aufschluss über den Einfluss der Länge auf die Platzierung in den Charts geben. Dazu wurde ein Programm geschrieben, welches die Liedlänge semiautomatisch ermittelt.

Die Basis des Programms ist eine CSV-Datei, welche zu jedem Lied die Lied-ID, den Interpret und den Titel enthält. Mithilfe dieser Daten startet das Programm eine Google-Videos-Suchanfrage. Die Zeitangabe des ersten Ergebnisses wird dann in Sekunden umgerechnet und als Standardwert eingetragen. An diesem Punkt ist die Korrektheit der Suche und des Ergebnisses manuell zu prüfen. Ist der angegebene Wert passend, kann dieser durch die Taste „Speichern“ übernommen werden, eine Anpassung der Zeitangabe ist dabei ebenso möglich. Liefert die Suche keine geeigneten Ergebnisse, kann die Taste „Unsicher“ betätigt werden, um vorerst den gleichnamigen Wert zu speichern und später eine genauere Prüfung vorzunehmen. Ist eine dieser Aktionen erfolgt, wird das nächste Lied geladen. Das Programm ist dem Anhang zu entnehmen. Ergebnis dieses Vorgangs ist ebenfalls eine CSV-Datei, welche zusätzlich zu den bislang enthaltenen Daten die Dauer in Sekunden und eine Bearbeitungs-ID enthält. Diese kann danach in KNIME eingelesen werden und mithilfe von einem Joiner-Knoten über die Lied-ID mit dem ursprünglichen Lied sowie allen weiteren Daten verknüpft werden.

4.2 Preprocessing

Beim Preprocessing geht es um die Aufbereitung der Datenbasis, um in darauffolgenden Schritten auf eine einheitliche, konsistente Grundlage zurückgreifen zu können. Dazu werden in der Regel mehrere Techniken verwendet, die im Folgenden am Workflow zur Vorbereitung der Liedtexte erklärt werden. Das Ziel ist es, die Datenbasis auf das Wesentliche zu konzentrieren und dadurch sowohl die Effektivität als auch die Effizienz des Modells zu verbessern [14].

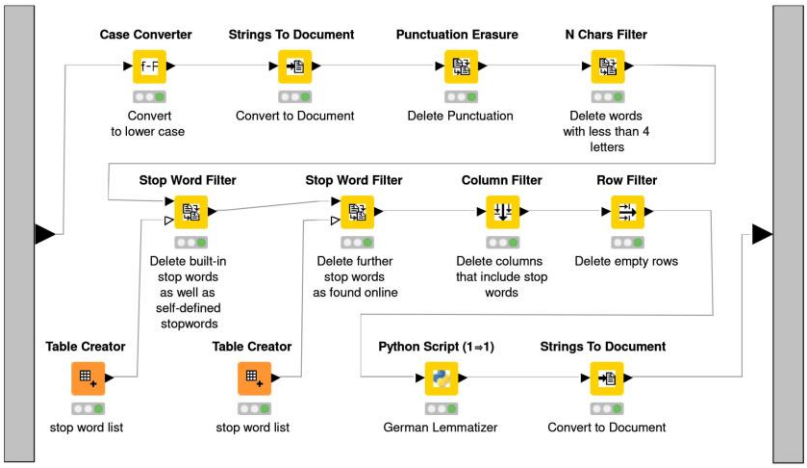

Abbildung 2. Workflow zum Preprocessing der Datenbasis

Sind die Daten, bestehend aus Lied-ID und Liedtext, durch eine CSV-Datei in KNIME eingelesen, startet das Preprocessing (siehe Abbildung 2). Im ersten Schritt werden alle Zeichen der Songtexte zu Kleinbuchstaben konvertiert, damit im Nachhinein keine Unterscheidung zwischen Groß- und Kleinschreibung erforderlich ist. Der dabei entstehende Informationsverlust ist durch die Vereinfachung der weiteren Analysen begründbar. Nachdem die Texte zu einem Dokument umgewandelt wurden, können nun sowohl einzelne Symbole als auch Wörter mit weniger als vier Zeichen entfernt werden. Diese Grenze wurde gewählt, da viele der Wörter mit weniger Buchstaben oftmals keine Bedeutung haben, die im weiteren Vorgehen von Interesse wäre. Beispiele hierfür sind Artikel, Pronomen, aber auch Füllwörter. Ausnahmen bestehen, werden aber durch die maßgebliche Simplifizierung der Datenbasis in Kauf genommen. Wäre die Grenze höher gewählt wurden, würde die Gefahr, Wörter mit einer entscheidenden Bedeutung wie „Mann“ oder „Frau“ zu entfernen, steigen. Des Weiteren werden mithilfe von unterschiedlichen Listen Stoppwörter entfernt. Nachdem anfänglich nur die im Wahlpflichtmodul gemeinsam erarbeitete Liste in Verbindung mit der von KNIME bereitgestellten Stoppwortliste für die deutsche Sprache genutzt wurde, wurde schnell klar, dass diese nicht alle notwendigen Stoppwörter enthielten. Aus diesem Grund wurde eine zweite, ergänzende Liste hinzugefügt [15]. Beim Entfernen der Stoppwörter wurden neue Spalten angelegt, die den transformierten Liedtext enthalten. Durch den Column-Filter-Knoten werden die Spalten, in denen noch Stoppwörter vorhanden sind, entfernt. Außerdem werden leere Datensätze, falls durch die vorherigen Schritte entstanden, entfernt. Anschließend folgt die Umformung jedes Wortes, um die verschiedenen Formen auf eine Grundform reduzieren zu können. Wie in Kapitel 2 empfohlen, wurden dazu anfänglich verschiedene Stemmer getestet, das beste Ergebnis konnte jedoch mit dem abgebildeten Lemmatizer erzielt werden. KNIME enthält zwar standardmäßig einen Lemmatizer, dieser ist allerdings auf die englische Sprache ausgerichtet. Deshalb wurde ein Python Skript geschrieben, welches die Überführung der Wörter in ihre Grundform, das Lemma, übernimmt (siehe

Anhang). Dieses iteriert zeilenweise über die Texte und formt jedes Wort in sein Lemma um, welches die folgende Themenanalyse vereinfacht, da eine geringere Wortvarianz erreicht wird, die Semantik aber vorhanden bleibt. Abschließend werden die Liedtexte erneut zu einem Dokument konvertiert und zusammen mit der Lied-ID in die weitere Bearbeitung gegeben.

5 Feature Engineering

5.1 Themenanalyse

Die relevantesten Features der Analyse sind die in Liedtexten zugrundeliegenden Themen. Hierfür wurden in der Methodologie zwei größere Teilschritte identifiziert: Zunächst wird LDA auf die bereinigten Liedtexte angewendet, um pro Lied 10 Themen zu identifizieren. Dies kann nicht der direkte Input für den Lernprozess des Entscheidungsbaumes darstellen, da dieser wahrscheinlich aufgrund der großen Menge an resultierenden Themen (7504) unüberprüfbar und invalide sein würde. Folglich ist es notwendig, die möglichen Thematiken durch ein Clustering zu reduzieren.

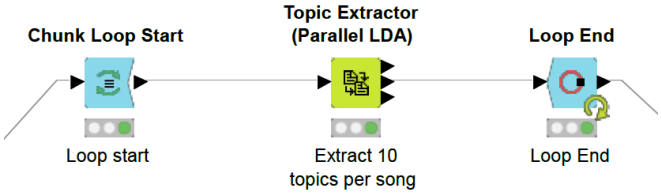

Abbildung 3. Topic Detection pro Liedtext

Für jedes Lied werden in einem Schleifenkonstrukt anhand des Topic-Extractor-Knotens 10 Themen extrahiert (siehe Abbildung 3). Diese Anzahl wurde gewählt, um ein Word Embedding auf eine größere Datenbasis anwenden zu können, welche jedem Wort einen Vektorraum mit 300 Dimensionen zuordnet.

Die zugrundeliegenden Vektoren wurden auf Grundlage deutscher Begriffe trainiert und stammen vom Unternehmen Deepset [16]. Zur Wort-Repräsentation verwendet es GloVe, welches Wikipedia als Datengrundlage benutzt. Die Vektoren werden anschließend aufgespalten und die daraus resultierenden Daten werden in den Folgeschritten für das Clustering vorbereitet.

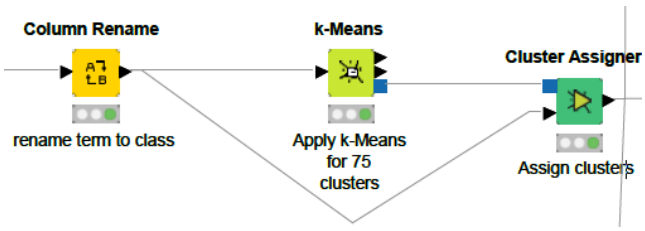

Abbildung 4. Clustering der Themen durch k-Means Algorithmus

Daraufhin folgt die Anwendung des Clustering auf Basis des k-Means Algorithmus, welcher die 300 Dimensionen aller Vektoren pro Thema verwendet, um Ähnlichkeiten und folglich potenzielle Themenfelder zu finden. Die 7504 Themen werden durch den Cluster-Assigner-Knoten in 75 Cluster automatisiert zugeordnet (siehe Abbildung 4). Es wurden mehrere unterschiedliche Clustergrößen getestet, die am besten zu interpretierenden Ergebnisse konnten durch die Anzahl 75 erzielt werden. Bei 100 Clustern waren diese zu granular, bei 50 Clustern waren einige zu groß. Somit wurde sich an die Anzahl 75 über mehrere Schritte angenähert.

Das Ergebnis des vergangenen Schrittes ist eine Einteilung der Themen in Cluster, die jedoch noch keine Namen besitzen (bisher Cluster 0, etc.). Hier ist eine manuelle und analytische Betrachtung der einzelnen Cluster notwendig, da keine der getesteten Algorithmen, wie z. B. k-Nearest Neighbour, zufriedenstellende Ergebnisse bei einer automatisierten Benennung liefern.

Die meisten Cluster konnten durch manuelle Betrachtung effizient benannt werden, z. B. „Emotionen“, „Körper“, „Bildung“ etc. Einige Cluster bieten keine menschlich erkennbaren Muster oder besitzen eine nicht handhabbare Größe.

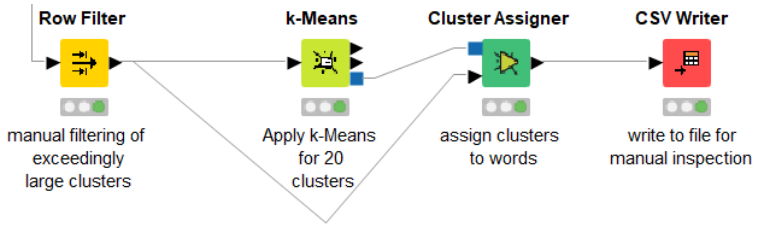

Abbildung 5. Manuelle Clusterbenennung

In der Folge wird im in Abbildung 5 gezeigten Schritt ein weiteres Teilclustering durchführt. Es wird ein „Problemcluster“ per Row-Filter-Knoten identifiziert und nochmals manuell per Clustering untersucht. Eine ebenfalls angewandte Alternative für „Problemfälle“ besteht in der Berechnung eines Median-Vektors für ein gegebenes Cluster, der wieder mit dem Word Embedding Modell verglichen wird. Die dem Median-Vektor ähnlichsten Begriffe geben ebenfalls Aufschluss über mögliche Namenskandidaten. In diesem Schritt ist aufgefallen, dass die Datenbereinigung im Preprocessing noch verfeinert werden könnte, da eine Vielzahl von Wörtern für die Analyse bedeutungslos ist und deshalb als „Undefiniert“ klassifiziert wurde.

Folgende Themenbereiche konnten identifiziert werden: Alltag, Architektur, Auto, Beziehungen, Bildung, Deutschland, Deutschrap, Emotionen, Englisch, Erfolg/Niederlage, Familie, Farben und Symbolik, Finanzen, Frauennamen, Geographie, Geschichte/Mythologie, Körper, Krieg, lateinische Sprachen, Lebensmittel, Medien, Musik, Musikindustrie, Osten, Räuber/Gendarm, Religion, Sport, Statistik, Technik, Tiere, Undefiniert, Verbrechen, Verkehr, Weltall und Wettbewerb.

Die gefundenen Clusterbezeichnungen wurden in ein Dictionary übertragen (Key: Clusternummer, Value: Clusterbezeichnung). Dieses wurde auf den Output von Abbildung 4 abgebildet und in eine CSV-Datei geschrieben. Das Ergebnis dieses

Workflows ist eine surjektive Abbildung aller möglichen 7504 Themen der untersuchten Lieder auf das entsprechende Cluster. Dies ermöglicht einen Rückschluss für jedes Lied auf ein oder mehrere Cluster, was für den kommenden Entscheidungsbaum einen verständlichen Input darstellt.

5.2 Wörteranzahl

Ein wichtiges Feature kann auch die Anzahl der Wörter in einem Text darstellen. Um diese mit einzubeziehen, wurden zwei verschiedene Herangehensweisen gewählt.

Zuerst wurde pro Lied die gesamte Wörteranzahl ohne Stoppwörter gezählt und diese in eine Liste geschrieben. Außerdem wurde neben der gesamten Wörteranzahl auch die Anzahl der Wörter ohne Duplikate gezählt, um feststellen zu können, wie viele unterschiedliche Wörter in dem Text verwendet wurden. Die Daten wurden anhand von SQL-Abfragen aus den vorher ermittelten Texten erhoben (siehe Anhang).

5.3 Jahrzehnte

Die zuvor genannten Features wurden vorbereitet, um die Gesamtheit der in der Nordakademie-Datenbank befindlichen deutschen Lieder zu verwenden. Folglich könnte ein Entscheidungsbaum für die Hit-Klassifikation des Gesamtzeitraums von über 40 Jahren angefertigt werden. Dennoch ist dies aufgrund geschichtlicher und zeitlicher Einflüsse auf die Liedthemen nur wenig sinnvoll, weshalb Lieder in Zeitperioden untersucht werden.

Zunächst wurden einzelne Jahrfünfte betrachtet. Dies führt zu einer zu geringen Lerngrundlage für den Entscheidungsbaum, weshalb der Betrachtungszeitraum auf die jeweiligen Jahrzehnte zwischen 1980 und 2020 erweitert wurde.

6 Durchführung des Experimentes

In diesem Schritt werden alle zuvor ermittelten Features aus einem Jahrzehnt pro Lied zusammengefasst. Die Ergebnisse hieraus werden in einen Workflow bei KNIME über mehrere Joiner-Knoten zusammengefügt. Hieraus entsteht die Datenbasis, die folgende Charakteristika über die Lieder enthält: Lied-ID, Themen 1-10, Lied-Dauer in Sekunden, Wortvarianz und Gesamtwörteranzahl sowie zusätzlich die Klassifikation Hit oder Nicht-Hit (siehe Abbildung 6). Die Klassifikation wurde im Voraus anhand der Definition aus Kapitel 3 mittels SQL-Abfragen vorgenommen.

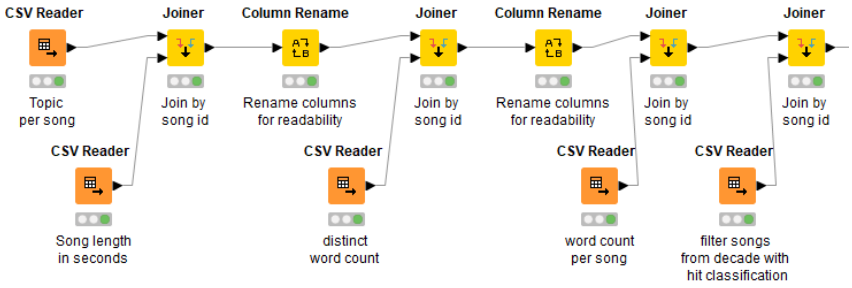

Abbildung 6. Zusammenführung der Features

Im nächsten Schritt des Experiments werden nun die vorher ermittelten Cluster zu den Themen pro Lied zugeordnet. Hier wurde sich dafür entschieden, die drei häufigsten Themencluster pro Lied zu erhalten und diese als Thema1, Thema2 und Thema3 zu bezeichnen. Dieses Vorgehen wurde gewählt, um die wichtigsten Cluster zu wählen und um dem Entscheidungsbaum nicht zu viele Entscheidungsmöglichkeiten zu geben.

Diese Daten werden in einem weiteren Schritt nun partitioniert, 85% der Daten sind hierbei die Daten, mit denen der Entscheidungsbaum die Klassifizierung lernt. Mit den restlichen 15% der Daten wird die Genauigkeit des Entscheidungsbaums validiert, indem diese Daten in den vorher erlernten Entscheidungsbaum eingegeben werden, der über den Hit-Status entscheiden soll. Die ermittelten Hit-Werte werden dann den tatsächlichen Werten gegenübergestellt und die Genauigkeit kann anhand einer Konfusionsmatrix berechnet werden.

7 Ergebnisse

Die untenstehenden Tabellen wurden anhand der Konfusionsmatrizen der Jahrzehnte erstellt. Die Testdaten wurden genutzt, um die Genauigkeit (Accuracy), die Fehlerquote (Error rate), der positive Vorhersagewert (Precision) sowie die Trefferquote (Recall) des vorher errechneten Entscheidungsbaums zu ermitteln. Für den ersten Knoten wurde Thema1 gewählt, da es sich hierbei um das für diese Arbeit relevanteste Feature handelt. Die Konfusionsmatrizen sind dem Anhang zu entnehmen.

Tabelle 1. Klassifikationsmetriken zur Hit-Klassifikation

Jahrzehnt	Datensätze Gesamt	Anzahl Testdaten	Accuracy	Error rate	Precision	Recall
1980er	461	53	0,566	0,434	0,318	0,467
1990er	463	48	0,708	0,292	0,273	0,333
2000er	1081	131	0,687	0,313	0,333	0,323
2010er	1540	186	0,710	0,290	0,292	0,159
Alle Jahrzehnte	3545	418	0,668	0,332	0,304	0,321

Tabelle 2. Klassifikationsmetriken zur Nicht-Hit-Klassifikation

Jahrzehnt	Datensätze Gesamt	Anzahl Testdaten	Accuracy	Error rate	Precision	Recall
1980er	461	53	0,566	0,434	0,742	0,605
1990er	463	48	0,708	0,292	0,838	0,795
2000er	1081	131	0,687	0,313	0,792	0,800
2010er	1540	186	0,710	0,290	0,772	0,880
Alle Jahrzehnte	3545	418	0,668	0,332	0,786	0,770

8 Interpretation

Die in Kapitel 7 vorgestellten Ergebnisse zeigen, dass das entwickelte Modell nur bedingt aussagekräftig ist. Dies lässt sich vor allem aus den globalen Werten aller Jahrzehnte schließen. Die in Tabelle 1 aufgezeigte Accuracy des Entscheidungsbaumes zeigt, dass im Schnitt 67 von 100 Liedern korrekt klassifiziert werden. Hier ist anzumerken, dass ein hoher Accuracy-Wert nicht notwendigerweise ein gut funktionierendes Modell beschreibt. Im Beispiel der Annahme, dass bei 1000 Liedern nur 10 Lieder Hits darstellen, und pauschal alle 1000 Lieder als Nicht-Hits klassifiziert werden, so hätte der Entscheidungsbaum eine Accuracy von 99%, dennoch verfehlt das Modell das Ziel, einen Hit zu klassifizieren. Analog hierzu ist die Error Rate zu werten.

Aus diesem Grunde wurden weitere Metriken zur Bewertung des Modells herangezogen, die Precision und der Recall. Die erstgenannte Metrik beschreibt den Anteil der als „richtig Positiv“ klassifizierten Testdaten im Verhältnis zu allen vorhergesagten Hits (richtig Positiv + falsch Positiv). Hingegen beschreibt der Recall den Anteil der als „richtig Positiv“ klassifizierten Hits im Verhältnis zu allen tatsächlichen Hits (richtig Positiv + falsch Negativ). An beiden Metriken ist zu erkennen, dass das Modell des vorgestellten Entscheidungsbaumes verbesserungswürdig ist. Über alle Jahrzehnte wird in der Precision ein Wert von 30% und ein Recall von 32% erreicht (siehe Tabelle 1). Dies bedeutet, dass von 100 vorhergesagten Hits nur 30 tatsächlich Hits waren und dass von 100 tatsächlichen Hits nur 32 Lieder als solche klassifiziert wurden. Zwar liegt eine negative Korrelation zwischen dem Verhältnis von Hit zu Nicht-Hit und der steigenden Anzahl der Lieder vor, ein Zusammenhang zur Precision/Recall konnte jedoch nicht festgestellt werden.

Das beste Ergebnis in Bezug auf Precision und Recall konnte hierbei in den Jahrzehnten 1980 und 2000 erzielt werden. Die größere Anzahl an Datensätzen kann anhand eines deutlich höheren Anteils deutscher Lieder in den Charts ab 2000 begründet werden. Ob ein Zusammenhang zwischen der Anzahl von Datensätzen und Recall/Precision existiert, konnte hier nicht ausreichend festgestellt werden. Es ist jedoch anzunehmen, dass diese Werte unabhängig voneinander sind (siehe Tabelle 1).

Als Grund für die nur bedingte Aussagekraft der dargestellten Ergebnisse ist der geringe Anteil von Hits am Gesamtdatensatz zu nennen, wobei die relativ große Anzahl an möglichen Themen das Ergebnis weiterhin negativ beeinflusst. Zudem sind viele der vorher herausgefundenen Themen als „Undefiniert“ klassifiziert, weshalb diese keine

Aussagekraft im Entscheidungsbaum tragen. Um die Aussagekraft des Entscheidungsbaumes zu verbessern, ist ein genaueres Preprocessing zu wählen, bei denen inhaltslose Wörter im Vorfeld aus den Daten bereinigt werden. In der vorliegenden Arbeit lag der Fokus auf vorgefertigten Stoppwortlisten. Diese sollten jedoch auf die Anforderungen des jeweiligen Projektes angepasst werden. Aus dem Korpus der durch LDA analysierten Themen wurden ca. 40% als „Undefiniert" gekennzeichnet. Folglich hätte nahezu die Hälfte der Themen als Stoppworte gelten können, wodurch ein besseres Ergebnis denkbar wäre. Des Weiteren hätte eine genauere Untersuchung zur Prüfung der Eignung von LDA stattfinden müssen. Dieser Ansatz wurde anhand des im Kapitel 2 beschriebenen Stand der Forschung gewählt, jedoch wird von einigen Autoren auf Einschränkungen dieses Wahrscheinlichkeitsmodells hingewiesen, die im Laufe dieser Arbeit weniger beleuchtet wurden, da der Fokus der Anwendung auf einer State-of-the-Art-Lösung lag.

Trotz der vorgestellten Limitierungen des Entscheidungsbaumes zur Klassifizierung von Hits, ist dieser hinsichtlich der Klassifizierung von Nicht-Hits dennoch nutzbar. Dies ist an der Tabelle 2 klar erkennbar und dadurch zu begründen, dass für die Nicht-Hits eine deutlich größere Lerngrundlage vorliegt. Dies bedeutet, dass von 100 vorhergesagten Nicht-Hits 79 tatsächlich Nicht-Hits waren und dass von 100 tatsächlichen Nicht-Hits 77 Lieder als solche klassifiziert wurden.

Daraus lassen sich folgende Aussagen über die Nicht-Hits der Jahrzehnte treffen:

In den 1980er-Jahren führte eine Konzentration auf das Thema Religion zu 60% dazu, dass das Lied nicht mindestens zwei Wochen in den Top 20 der Charts bleiben konnte. Ähnlich sieht es bei den Themen Bildung (82%), Geschichte/Mythologie (76%) und Beziehungen (69%) aus. Wird ein erster Knotensplitt nicht auf das Thema1 forciert, ergibt sich für den ersten Knoten eine Aufspaltung nach der Wortvarianz. In 84% der Fälle führt eine Wortvarianz von unter 86 Wörtern zu einem Nicht-Hit.

In den 1990er-Jahren zeichnet sich ein ähnliches Bild ab. Hier führen die Themen Religion (85%), Geschichte/Mythologie (86%) und Beziehungen (73%) zu einem Misserfolg. Zusätzlich hierzu ist zu erkennen, dass das Thema Geographie im Gegensatz zu den 1980er Jahren zwar weniger häufig thematisiert wurde, allerdings auch häufiger zum Nicht-Hit führte (89%). Das Thema Bildung hingegen hat an Bedeutung gewonnen, da in diesem Jahrzehnt nur noch 63% zu einem Nicht-Hit führten. Selbst bei Herausnahme des voreingestellten Wurzelknotens berechnet der Entscheidungsbaum trotzdem das Thema1 als wichtigstes Feature zur Identifizierung von Hit/Nicht-Hit.

Auch in den 2000er-Jahren ist zu erkennen, dass eine Thematisierung von Religion (79%), Geschichte/Mythologie (77%), Beziehungen (80%), Geographie (78%) und Bildung (74%) mit hohen Wahrscheinlichkeiten zu Nicht-Hits führen. Zusätzlich hierzu sind die Themen Wettbewerb (78%) und Verkehr (74%) mögliche Indikatoren für Nicht-Hits. Wird der voreingestellte Knotensplit entfernt, ist die Wortvarianz von unter 82 Wörtern zu 92% das ausschlaggebende Feature von Nicht-Hits.

In den 2010er-Jahren kann ein teilweiser Themenwandel in Bezug auf die Nicht-Hits festgestellt werden. Obwohl die Themen Religion (75%), Beziehungen (82%), Geographie (76%) und Verkehr (79%) nach wie vor meistens zu Nicht-Hits führen, lassen sich zusätzlich deutsche Lieder mit englischem Einfluss als einer der wichtigsten

Indikatoren für einen Nicht-Hit mit 79% identifizieren. Geschichte/Mythologie und Bildung wird in diesem Jahrzehnt hingegen weniger thematisiert. Wird der Knotensplit nicht auf Thema1 forciert, ist weiterhin die Wortvarianz das ausschlaggebende Feature, wobei dieses in diesem Fall irrelevant ist, da die Lieder in beiden Ästen des Splits größtenteils als Nicht-Hits klassifiziert werden.

9 Fazit

Gegenstand der Untersuchung war das Identifizieren von klassifizierenden Merkmalen und anhand dessen die Klassifikation von Liedern als Hit anhand eines Entscheidungsbaumes. Die Methoden und benutzten Features wurden auf den in Kapitel 2 identifizierten State-of-the-Art definiert. Um dieses Experiment durchzuführen, wurde sich am DASC-PM orientiert. Die Daten wurden im Rahmen des Prozessschrittes Datenbereitstellung nach der Erhebung betrachtet, um diese für das Preprocessing vorzubereiten. Anschließend wurden die Features extrahiert, um einen Entscheidungsbaum für den Prozessschritt Analyse erstellen zu können.

Das ursprüngliche Ziel, ein Lied als Hit klassifizieren zu können, konnte nur bedingt erreicht werden. Hingegen ist das Modell in der Lage, die Merkmale von Nicht-Hits zu identifizieren. Dies spiegelt sich in vergleichsweise hohen Werten für Precision und Recall wider. Folglich konnten unter Anderem Themen identifiziert werden, die über die Jahrzehnte hinweg mögliche Indikatoren für Nicht-Hits darstellen, wie z.B. Religion und Beziehungen. Zudem wurden Themen identifiziert, die für bestimmte Jahrzehnte spezifisch sind. So sind deutsche Lieder mit englischem Einfluss vor allem in den 2010er-Jahren prävalent zur Identifikation von Nicht-Hits. Die Wortvarianz scheint ebenfalls ein maßgebliches Feature zur Beschreibung von Nicht-Hits zu sein. Hingegen sind die Wortanzahl und die Liedlänge im beschriebenen Modell keine ausschlaggebenden Features.

Abschließend lässt sich somit sagen, dass ein Modell zur Identifikation von Hits weiteren Forschungsbedarf erfordert und die gegebenen Features hierfür nicht ausreichen. Andere Features, die in Kapitel 2 identifiziert wurden, wie z.B. die Melodie oder der Popularitätsgrad der/des Interpreten, könnten die Güte für eine Hit-Klassifikation erhöhen. Der gegebene Rahmen der Untersuchung war hingegen nicht ausreichend, um weitere Features zu erheben.

Literatur

[1] E. Zangerle et al., „Hit Song Prediction: Leveraging Low- and High-Level Audio Features“, *Proceedings of the 20th International Society for Music Information Retrieval Conference 2019 (ISMIR 2019)*, S. 319–326, 2019. [Online]. Verfügbar unter: https://archives.ismir.net/ismir2019/paper/000037.pdf

[2] M. Fell und C.Sporleder, „Lyrics-based Analysis and Classification of Music“ in *Proceedings of COLING 2014*, S. 620–631. [Online]. Verfügbar unter: https://www.aclweb.org/anthology/C14-1059

[3] F. Pachet und P. Roy, „Hit Song Science Is Not Yet a Science“, *Proceedings of the 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA*, S. 355–360, 2008. [Online]. Verfügbar unter: https://www.cs.swarthmore.edu/~turnbull/cs97/f08/paper/pachet08.pdf

[4] Y. Kim, B. Suh und K. Lee, „#nowplaying the future billboard: Mining Music Listening Behaviors of Twitter Users for Hit Song Prediction“ in *Proceedings of SoMeRA'14*, S. 51–56, doi: 10.1145/2632188.2632206.

[5] D. M. Blei, A. Y. Ng und M. I. Jordan, „Latent Dirichlet Allocation“, *J. Mach. Learn. Res.*, Jg. 3, null, S. 993–1022, 2003.

[6] N. Teneva und W. Cheng, „Salience Rank: Efficient Keyphrase Extraction with Topic Modeling“ in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, S. 530–535, doi: 10.18653/v1/P17-2084.

[7] Y. Suzuki und F. Fukumoto, „Detection of Topic and its Extrinsic Evaluation Through Multi-Document Summarization“ in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, S. 241–246, doi: 10.3115/v1/P14-2040.

[8] R. Rianto, A. B. Mutiara, E. P. Wibowo und P. I. Santosa, *Improving the Accuracy of Text Classification using Stemming Method, A Case of Informal Indonesian Conversation*, 2020.

[9] F. Boudin, H. Mougard und D.Cram, „How Document Pre-processing affects Keyphrase Extraction Performance“, *Proceedings of the 2nd Workshop on Noisy User-generated Text*, S. 121–128, 2016. [Online]. Verfügbar unter: https://www.aclweb.org/anthology/W16-3917.pdf

[10] K. Grąbczewski, *Meta-Learning in Decision Tree Induction*. Cham, SWITZERLAND: Springer International Publishing AG, 2013. [Online]. Verfügbar unter: http://ebookcentral.proquest.com/lib/nordakademie/detail.action?docID=3091992

[11] T. Boros, S. D. Dumitrescu und S. Pipa, „Fast and Accurate Decision Trees for Natural Language Processing Tasks“ in *Proceedings of Recent Advances in Natural Language Processing*, S. 103–110, doi: 10.26615/978-954-452-049-6_016.

[12] *CRISP-DM 1.0*, CRISP-DM Consortium, 2000. [Online]. Verfügbar unter: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

[13] M. Schulz *et al.*, *DASC-PM v1.0: Ein Vorgehensmodell für Data-Science-Projekte*. Elmshorn: Nordakademie, Hochschule der Wirtschaft, 2020.

[14] V. Gurusamy und S. Kannan, Hg., *Preprocessing Techniques for Text Mining*. Madurai Kamaraj University: Madurai Kamaraj University, 2015.

[15] T. Ehling, „Deutsche Stoppwort-Liste » tim-ehling.com“, *tim-ehling.com*, 2016, 2016-08-11CEST20:10:44+02:00. [Online]. Verfügbar unter: https://tim-ehling.com/deutsche-stoppwort-liste-400/. Zugriff am: 13. Dezember 2020.

[16] *German Word Embeddings*. DeepSet GmbH. [Online]. Verfügbar unter: https://deepset.ai/german-word-embeddings