HAW HAMBURG

Jelte Dreetz, Mareike Marx, Luca Schmitt - im Rahmen des Projekt Datascience

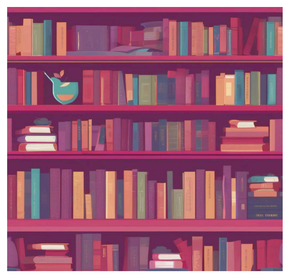
Ist Harry Potter ein Kochbuch?

Die Klassifikation von Buchtitel nach Genre ist entscheidend, um Inhalte korrekte einzuordnen und Fehlinterpretationen, wie die Frage "Ist Harry Potter ein Kochbuch?" zu vermeiden. Unser Projekt zeigt, wie maschinelles Lernen genutzt werden kann, um Titel präzise Genres präzise zuzordnen. Zusätzlich untersuchen wir die Beziehung zwischen Buchtitel und Bewertungen, um Einblicke in Leserpräferenzen zu gewinnen.

Datensets

Als Grundlage für das trainieren wurden die folgenden Datensets verwendet.
Bei Goodreads handelt es sich um einen Sammlung von Tabellen.

[1]: Goodreads [2]: Amazon Bestselling [3]: Amazon Trending [4]: wonderbk



Rating Modell

Fragestellung: kann man anhand des Buchtitels das Rating feststellen?

I. Methode

Es wurden Datensätze über Kaggle gesammelt[1] und Goodreads Datensätze wurden zusammengefügt und bereinigt.
Folgende Datenzeilen wurden entfernt:

- mindestens einen "null", "none", oder "nan" Wert
- Wenn der Buchtitel ein leerer String ist
- Duplikate
- Wenn Buchtitel Sonderzeichen enthalten (Es werden nur ASCI Zeichen akzeptiert)
- Wenn das Buch kein Rating hat

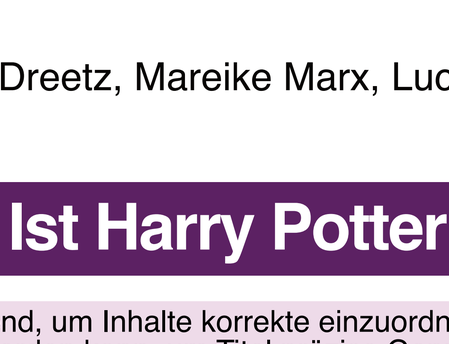


Abb.1: Verteilung der Ratings in Trainings- und Testdatensatz

Nach der Bereinigng hat der Datensatz 1.276.954 Einträge. (Hinweis: Für das Training wurde nicht der gesamte Datensatz verwendet, weil die Rechenleistung dafür nicht ausgereicht hat.)

Die Ratings wurden von Gleitkommazahlen in ganze Zahlen umgewandelt.

Es wurde ein DistilBERT Model mit einem gesäuberten Rating-Datensatz trainiert. Es wurden 100000 Zeilen des Datensatzes zum Trainieren und Testen verwendet (80000 für Training, 20000 für Test).

Der Datensatz ist normalverteilt. Damit die Verteilung in Trainings- und Testdatensatz gleich ist, wurde eine stratifizierende Methode verwendet, um den Datensatz aufzuteilen.

II. Ergebnisse

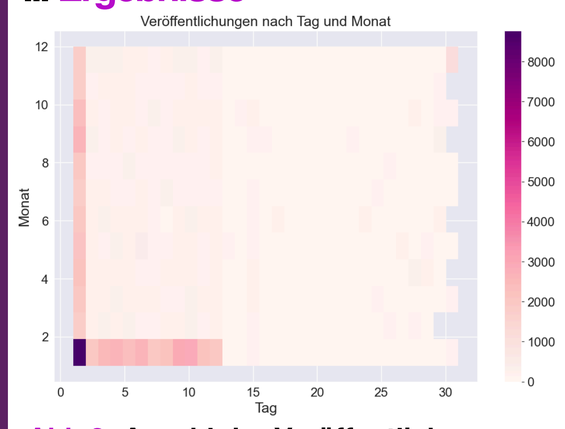


Abb.2: Anzahl der Veröffentlichungen nach Tag und Monat

➡ 01.01 auffällig oft als Veröffentlichungsdatum angegeben

Schlussfolgerung: Es ist möglich, dass der 01.01 als Platzhalter verwendet wurde, wenn das tatsächliche Erscheinungsdatum unbekannt oder nicht verfügbar war.

Nach dem Herausfiltern der Einträge mit dem 01.01 ergibt sich eine gleichmäßigere Verteilung der Veröffentlichungsdaten, wobei der 31.12 häufiger als Veröffentlichungsdatum erscheint.

Dies könnte darauf hinweisen, dass das Jahresende ebenfalls als Standard- oder Platzhalterdatum ausgewählt worden sein könnte, wenn kein genaues Datum bekannt war.

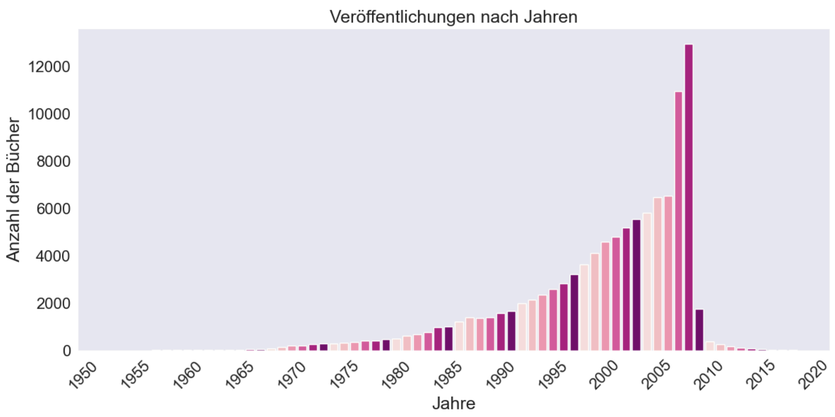


Abb.3: Anzahl der Veröffentlichungen nach Jahren

➡ auffällig starker Rückgang der veröffentlichten Titel ab dem Jahr 2009

Schlussfolgerung:

- Datensätze nicht vollständig aktualisiert
- Beinhalten nicht alle neuen Buchveröffentlichungen nach 2009
- Fehlerhafte Datensätze (z. B. falsche Jahreszahlen)

Das Modell hat ein massives Underfitting Problem. Die Loss Werte halten sich über das gesamte Training bei ca. 90 %. Dies ist allerdings nicht verwunderlich. Es wäre naiv davon auszugehen, dass es einen direkten Zusammenhang zwischen Buchtitel und Buchbewertung gibt.

Tabelle 1: Ergebnisse der Rating Klassifikation mit DistilBERT

Rating	precision	recall	f1-score	support
1	0.00	0.00	0.00	52
2	0.00	0.00	0.00	474
3	0.55	1.00	0.71	10915
4	0.00	0.00	0.00	8067
5	0.00	0.00	0.00	492
accuracy			0.55	20000
macro avg	0.11	0.20	0.14	20000
weighted avg	0.30	0.55	0.39	20000

Achtung: Das Modell hat ausschließlich Rating 3 vorhergesagt. Der F1-Score ist schwach

Fazit: Es ist nicht möglich eine Buchbewertung basierend auf dem Titel vorherzusagen.

Interessante Beobachtung

Wenn das Modell nur, mit 10k Datenzeilen trainiert wird kann es mindestens 2 Ratings vorhersagen.

Das Modell, welches mit 100k Datenzeilen trainiert wurde kann nur ein Rating vorhersagen. Und zwar das Rating, welches am häufigsten vorkommt.

These: Das Modell findet keinen Zusammenhang zwischen Rating und Buchtitel und versucht daher das häufigste Rating vorherzusagen. Weil somit ist es am wahrscheinlichsten dass die Vorhersage korrekt ist.

Genre Modell

Fragestellung: kann man anhand des Buchtitels das Genres feststellen?

I. Methode

Datensätze wurden über Kaggle gesammelt. Datensätze von Amazon Bestseller wurden zusammengefügt und bereinigt (doppelte/ falsche entfernt).

Größe des Datensatzes [2,3,4]

- Trainingsdaten: 63.065 Titel ➡ Testdaten: 15.767 Titel
- Merkmale pro Datensatz: 5000 (TF-IDF-Vektoren oder gepaddete Sequenzen von Tokens)

Datenvorverarbeitung

- Tokenisierung: Die Buchtitel wurden in Sequenzen von Tokens (Wörtern) umgewandelt.
- Padding: Alle Sequenzen wurden auf eine einheitliche Länge von 100 Tokens gebracht, um die Verarbeitung in neuronalen Netzen wie RNNs zu erleichtern.
- TF-IDF-Vektorisierung: Für Random Forest und Naive Bayes wurden die 5000 häufigsten Wörter extrahiert und als Vektoren mit einer festen Länge von 5000 dargestellt.

Label-Vorbereitung

- Genres wurden mittels One-Hot-Encoding in binäre Vektoren umgewandelt, sodass jedes Genre durch einen eindeutigen Vektor repräsentiert ist.

Datenaufteilung

- Der Datensatz wurde mit einer stratifizierten 80/20-Aufteilung in Trainings- und Testdaten geteilt, um einegleichmäßige Verteilung der Genres sicherzustellen.

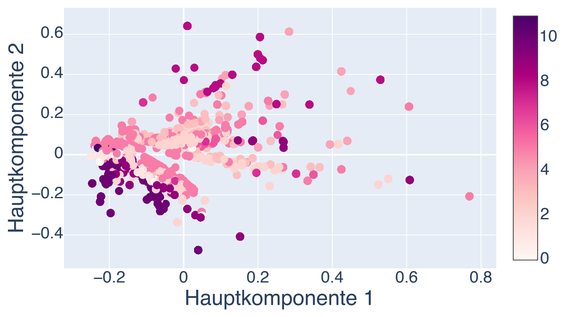
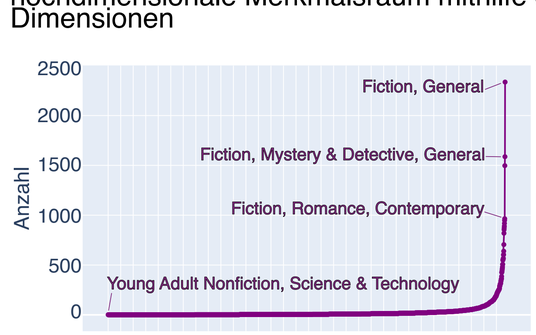
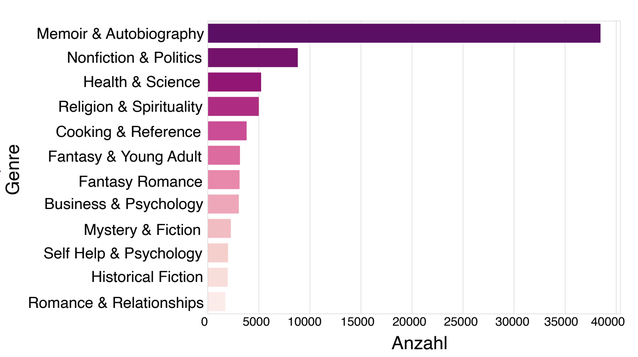


Abb.4: Streudiagramm der Genre Cluster. Visualisierung durch reduzieren des hochdimensionale Merkmalsraum mithilfe der Principal Component Analysis (PCA) auf zwei Dimensionen

- K Means-Algorithmus zur Gruppierung der Genres
- Die Elbow-Methode für Bestimmung der optimale Anzahl an Clustern. Dabei wurde die sogenannte Inertia berechnet, die die Summe der quadrierten Abstände aller Datenpunkte zu ihren zugeordneten Clusterzentroiden darstellt. Eine niedrige Inertia weist auf kompaktere und homogener gruppierte Cluster hin.
- Daten in 12 Cluster segmentiert und Zuordnung der Cluster zu Genregruppen wie z. B. „Cooking & Reference“ oder „Fantasy & Young Adult“





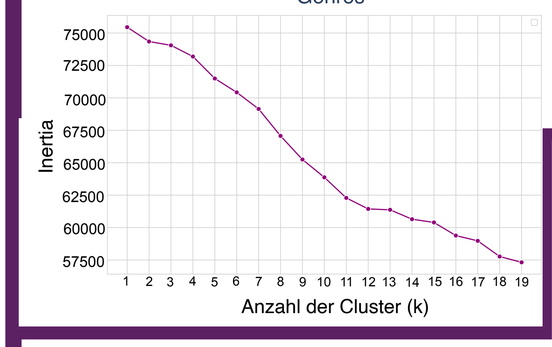


Abb.5: Häufigkeit aller Genres. Die K-Means-Clustering-Ergebnisse zeigen, dass sich Genres auf Basis ihrer textuellen Ähnlichkeit gruppieren lassen.

II. Ergebnisse

Klassifikationsverfahren	Accuracy
Naive Bayes	0,6229
Random Forest	0,6332
Künstliches Neuronales Netzwerk	0,6276
Convolutional Neural Network	0,6467
Gated Recurrent Unit Network	0,6299
DistilBERT	0,6932

Genre	Naive Bayes	Random Forest	KNN	CNN	GRU	DistilBERT
Business & Psychology	0,42	0,6	0,6	0,6	0,61	0,72
Cooking & Reference	0,81	0,82	0,81	0,83	0,81	0,87
Fantasy & Young Adult	0,13	0,27	0,31	0,33	0,05	0,41
Fantasy Romance	0,34	0,42	0,42	0,43	0,39	0,51
Health & Science	0,22	0,29	0,33	0,38	0,39	0,46
Historical Fiction	0,14	0,22	0,21	0,28	0	0,4
Memoir & Autobiography	0,72	0,73	0,72	0,74	0,73	0,77
Mystery & Fiction	0,31	0,49	0,47	0,47	0,45	0,58
Nonfiction & Politics	0,61	0,62	0,62	0,65	0,62	0,72
Religion & Spirituality	0,47	0,55	0,55	0,59	0,59	0,65
Romance & Relationships	0,09	0,25	0,37	0,34	0,2	0,47
Self Help & Psychology	0,38	0,48	0,53	0,58	0,57	0,62

Tabelle 3: Vergleich des F1-Score. Die Ergebnisse zeigen, dass DistilBERT die stärkste Leistung, in Bezug des F1-Scores, in allen Genregruppen erzielte. Darüberhinaus zeigt sich, dass Genregruppen mit klaren Schlüsselwörtern, wie bei "Cooking & Reference", einfacher zu klassifzieren sind, auch bei einer geringeren Datenmenge. Außerdem zeigt sicht, dass DistilBERT mti seinem vortrainiertem Wissen aus einem umfangreichenTextkorpora, besser in der Lage war, Titel bzw. Wörter, die fachspezifischer sind, zu den passenden Genres klassifizieren, wie bei "Historical Fiction".

Tabelle 2: Vergleich der Genauigkeit (Accuracy). DistilBERT erzielte mit **69,32%** die höchste Genauigkeit und übertraf damit die anderen Modelle. Random Forrest und Naive Bayes lieferten solide Ergebisse, insbesondere für Genres mit eindeutigen Schlüsselwörtern. Trotz ihrer vielversprechenden Leistung zeigten KNN und CNN eine deutliche Tendenz zum Overfitting, was auf die Notwendigkeit stärkerer Regularisierungsmaßnahmen hinweist. GRU-Modelle, die für die Verarbeitung von sequentieller Daten entwickelt wurden, boten eine robuste Leistung, blieben jedoch hinter DistilBERT zurück.[5]

Zusammenfassung

Bei der Genre Klassifikation bietet DistilBERT die beste Lesitung durch sein Verständnis von Kontext und Semantik. Traditionelle Algortihmen wie Naive Bayes und Random Forest sind bei der Verarbeitung komplexer Texte begrenzt, während neuronale Netze, wie ein CNN oder GRU, eine Brücke zwischen klassischen und modernen Ansätze schlagen. Bei der Rating Klassifikation war es uns nicht möglich einen Zusammenhang zwischen einem Buchtitel und seinem Rating zu finden. Daher vermuten wir, dass es höchstwahrscheinlich weitere Faktoren, wie der Autor oder der Inhalt des Buches, braucht.

[5] L. Tunstall, L. von Werra, T. Wolf, und A. Géron, *Natural language processing with transformers: building language applications with hugging face*, Revised edition. Beijing Boston Farnham Sebastopol Tokyo: O’Reilly, 2022.