Projekt: NLP DLBAIPNLP01_D

Projektbericht

Studiengang: Angewandte Künstliche Intelligenz

Sven Behrens

Matrikelnummer: 42303511

Tutor: Prof. Dr. Maja Popovic

15. Dezember 2025

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

API Application Programming Interface

1 Einleitung

Durch die zunehmende Digitalisierung werden Benutzerbewertungen sowohl zur Beurteilung von Produkten als auch in sozialen Medien immer wichtiger, sowohl für Kunden, um einen schnellen Überblick zu bekommen, als auch für Unternehmen, um Feedback systematisch auszuwerten. So werden täglich Millionen von Nutzermeinungen erzeugt, deren manuelle Analyse längst nicht mehr praktikabel ist. Die automatisierte Sentimentanalyse von Produktrezensionen stellt dabei eine zentrale Herausforderung im Bereich des Natural Language Processing dar, deren Bewältigung für Unternehmen maßgeblich zur Produktverbesserung und Kundenzufriedenheitsanalyse beiträgt. Vor diesem Hintergrund wurde im Rahmen des Moduls „Projekt: NLPän der IU Internationalen Hochschule ein umfassendes System zur Sentimentanalyse von Amazon-Produktbewertungen entwickelt, das sowohl klassische Machine-Learning-Verfahren als auch moderne Transformer-Architekturen systematisch vergleicht.

Das primäre Projektziel bestand in der Entwicklung und dem Vergleich verschiedener Klassifikationsansätze zur automatischen Sentiment-Erkennung aus englischsprachigen Produktrezensionen. Die zentrale Forschungsfrage konzentrierte sich darauf, wie sich traditionelle Machine-Learning-Verfahren, mit Modellen wie Logistic Regression gegenüber modernen vortrainierten Sprachmodellen wie DistilBERT hinsichtlich Klassifikationsgenauigkeit und Praxistauglichkeit verhalten.

Die Datenbasis bildeten Amazon-Produktrezensionen aus drei Kategorien, Automotive, Pet Supplies und Video Games, mit insgesamt 150.000 Datensätzen, die stratifiziert in Trainings- (70%), Validierungs- (15%) und Testdaten (15%) aufgeteilt wurden. Jede Rezension umfasst Titel, Text und eine Sternebewertung von 1 bis 5, wobei sowohl eine 5-Klassen-Klassifikation als auch eine aggregierte 3-Klassen-Sentimentanalyse (negativ, neutral, positiv) untersucht wurden.

Die methodische Vorgehensweise gliederte sich in mehrere aufeinander aufbauende Phasen. Nach einer initialen Datenaufbereitung mit Textbereinigung, Lemmatisierung und Stopwort-Entfernung erfolgte zunächst die Implementierung klassischer Modelle unter Verwendung von TF-IDF-Vektorisierung mit bis zu 20.000 Features. Hierbei wurden Logistic Regression, Naive Bayes, SVM, Random Forest und Gradient Boosting systematisch evaluiert. Parallel dazu wurde ein Fine-Tuning des DistilBERT-Modells auf dem vollständigen Trainingsdatensatz durchgeführt sowie ein vortrainiertes BERT-Modell für Sentimentanalyse im Zero-Shot-Modus getestet.

Der vorliegende Projektbericht dokumentiert systematisch den gesamten Entwicklungsprozess. Nach dieser Einleitung folgt die detaillierte Beschreibung der Methodik, einschließlich der Datenaufbereitung sowie der Modellauswahl und -konfiguration. Anschließend werden die Trainingsabläufe und deren Ergebnisse ausführlich dargestellt, wobei das beste Modell (DistilBERT) eine Testgenauigkeit von 80,91% erreichte, während die klassische Logistic Regression mit 76,66% konkurrenzfähige Referenzleistung erzielte. Abschließend werden die gewonnenen Erkenntnisse in einem Fazit zusammengefasst und kritisch reflektiert.

2 Hauptteil

2.1 Hardwareauswahl

Zu Projektbeginn standen drei Alternativen zur Verfügung: ein MacBook Pro mit Apple-M2-Chip, ein Windows-PC mit NVIDIA RTX 3060 sowie verschiedene Cloud-Computing-Anbieter.

Für das Training der klassischen Machine-Learning-Modelle wie Logistic Regression, Naive Bayes und SVM war sowohl das MacBook als auch der Windows-PC ausreichend, da diese Verfahren primär CPU-basiert arbeiten. Für das Fine-Tuning der BERT-Modelle schied das MacBook jedoch aufgrund der fehlenden CUDA-Unterstützung aus.

Das Training wurde daher auf dem Windows-11-Computer (Intel i7-12700K, 32 GB DDR5, NVIDIA RTX 3060 mit 12 GB VRAM) durchgeführt. Da selbst das längste Training (DistilBERT auf 105.000 Samples) mit etwa zwei Tagen in einem akzeptablen Rahmen blieb, waren zusätzliche Cloud-Lösungen nicht erforderlich.

2.2 Projektumgebung einrichten

Zu Beginn des Projekts wurde ein GitHub-Repository angelegt. Obwohl nur eine Person am Projekt arbeitet, ermöglicht GitHub eine nachvollziehbare Versionsverwaltung und ein einfaches Zurücksetzen auf frühere Stände. Anschließend wurde eine grundlegende Verzeichnisstruktur erstellt und mit `venv` eine virtuelle Umgebung eingerichtet, in der zentrale Bibliotheken installiert wurden.

2.3 Datenbeschaffung

Als Datenquelle diente der Amazon Reviews 2023 Datensatz Hou et al., 2024, der über 500 Millionen Produktbewertungen aus verschiedenen Kategorien umfasst und für akademische Forschungszwecke frei verfügbar ist.

Bei der Auswahl der Kategorien wurden zunächst kleinere Kategorien wie Gift Cards in Betracht gezogen. Diese erwiesen sich jedoch als ungeeignet, da die zugehörigen Rezensionen häufig nur aus sehr kurzen Texten wie „gutöder „schnelle Lieferung"bestanden, die für eine aussagekräftige Sentimentanalyse nicht ausreichend sind.

Stattdessen wurden die Kategorien Automotive, Pet Supplies und Video Games ausgewählt. Diese Kategorien zeichnen sich durch ausführlichere Rezensionen aus, in denen Nutzer ihre Erfahrungen detailliert beschreiben. Zudem repräsentieren sie unterschiedliche Produkttypen, wodurch die Generalisierungsfähigkeit der trainierten Modelle besser evaluiert werden kann.

Aus jeder Kategorie wurden jeweils 50.000 Rezensionen entnommen, sodass ein balancierter Gesamtda-

tensatz von 150.000 Datensätzen entstand.

2.4 Datenanalyse

2.4.1 Datensatzstruktur

In einem ersten Schritt wurden die Rohdaten untersucht und ihre Struktur analysiert. Die Rezensionen liegen im JSONL-Format vor, wobei jede Zeile einen JSON-Datensatz repräsentiert. Table 1 zeigt die verfügbaren Felder und ihre Bedeutung.

Tabelle 1: Struktur der Amazon-Rezensionsdaten im JSONL-Format

Feld	Beschreibung	Beispiel
rating	Sternebewertung (1.0–5.0)	5.0
title	Titel der Rezension	„Great product!"
text	Ausführlicher Rezensionstext	„Item came as described..."
images	Angehängte Bilder	[]
asin	Amazon Produkt-ID	B01LZA8SGZ
parent_asin	Übergeordnete Produkt-ID	B0BV88374L
user_id	Anonymisierte Benutzer-ID	AGXVBIUFLFGMV...
timestamp	Unix-Zeitstempel	1513092936205
helpful_vote	Anzahl hilfreicher Stimmen	0
verified_purchase	Verifizierter Kauf	true

Die Felder `title` und `text` wurden zu einem zusammenhängenden Eingabetext (X) kombiniert, während das Feld `rating` als Zielklasse (y) mit Werten von 1 bis 5 Sternen dient. Zusätzlich wurde die Produktkategorie gespeichert, um kategoriespezifische Evaluationen zu ermöglichen.

Anschließend erfolgte die Aufteilung der Daten in 70% Trainings-, 15% Validierungs- und 15% Testdaten.

2.5 Vorverarbeitungs-Pipeline

Die folgenden Ausführungen zur Textvorverarbeitung und Feature-Extraktion orientieren sich an den Grundlagen aus Schaaff (2025, S. 62–78).

Die Textvorverarbeitung bildet einen entscheidenden Schritt zur Vorbereitung der Rohtexte für die maschinelle Verarbeitung. Ziel ist es, Rauschen zu reduzieren und die relevanten sprachlichen Merkmale zu extrahieren.

2.5.1 Textbereinigung und Normalisierung

Die implementierte Pipeline verarbeitet jeden Rezensionstext in mehreren aufeinander aufbauenden Schritten:

1. **Textkombination**: Zusammenführung von `title` und `text` zu einem Eingabestring
2. **Kleinschreibung**: Konvertierung aller Zeichen zu Kleinbuchstaben zur Vereinheitlichung
3. **HTML-Bereinigung**: Entfernung von HTML-Tags mittels regulärer Ausdrücke
4. **URL-Entfernung**: Filterung von Weblinks, die keine semantische Relevanz besitzen
5. **Zahlenentfernung**: Eliminierung numerischer Werte
6. **Satzzeichenentfernung**: Entfernung aller Interpunktionszeichen
7. **Tokenisierung**: Zerlegung des Textes in einzelne Wörter mittels NLTK
8. **Stoppwort-Filterung**: Entfernung häufiger englischer Funktionswörter (the, is, at, ...)
9. **Lemmatisierung**: Rückführung der Wörter auf ihre Grundform mittels WordNetLemmatizer
10. **Längenfilterung**: Entfernung von Tokens mit weniger als 2 Zeichen

2.5.2 Feature-Extraktion

Für die Transformation der vorverarbeiteten Texte in numerische Vektoren stehen verschiedene Methoden zur Verfügung. Table 2 vergleicht die gängigsten Ansätze.

Tabelle 2: Vergleich der Feature-Extraktionsmethoden

Methode	Vorteile	Nachteile
Bag of Words	Einfach, schnell, interpretierbar	Ignoriert Worthäufigkeit im Korpus, große sparse Matrizen
TF-IDF	Gewichtet wichtige Wörter höher, reduziert Rauschen, bewährt für Klassifikation	Ignoriert Wortkontext und -reihenfolge
Word2Vec / GloVe	Erfasst semantische Ähnlichkeit, dichte Vektoren	Benötigt viele Trainingsdaten, schwerer interpretierbar

Für die klassischen Machine-Learning-Modelle wurde **TF-IDF** (Term Frequency-Inverse Document Frequency) gewählt. Diese Entscheidung basiert auf mehreren Faktoren: TF-IDF ist für Sentiment-Klassifikation bewährt, ermöglicht schnelles Training ohne GPU und bietet Interpretierbarkeit, es lässt sich nachvollziehen, welche Wörter zur Klassifikation beitragen.

Tabelle 3: Konfiguration der TF-IDF-Vektorisierung

Parameter	Wert	Beschreibung
max_features	10.000	Maximale Anzahl der Features
ngram_range	(1, 2)	Uni- und Bigramme
min_df	2	Term muss in mind. 2 Dokumenten vorkommen
max_df	0,95	Term darf in max. 95% der Dokumente vorkommen

Die Verwendung von Bigrammen ermöglicht die Erfassung von Wortpaaren wie „not good” oder „very bad”, die für die Sentimentanalyse besonders relevant sind, da sie Negationen und Verstärkungen berücksichtigen.

2.6 Modellauswahl

Um einen fundierten Vergleich zwischen klassische Machine-Learning-Verfahren und Transformer-basierte Sprachmodelle zu ermöglichen, wurden folgende Modelle evaluiert.

2.6.1 Klassische Machine-Learning-Modelle

Als Baseline dienten etablierte Klassifikationsverfahren in Kombination mit TF-IDF-Vektorisierung. Die Term Frequency-Inverse Document Frequency (TF-IDF) transformiert Texte in numerische Vektoren, wobei häufig vorkommende, aber wenig aussagekräftige Wörter herabgewichtet werden.

Folgende Modelle wurden implementiert und evaluiert:

- **Logistic Regression**: Bei der logistischen Regression wird eine S-förmige logistische Funktion (Sigmoidfunktion) an die Daten gefittet, wobei der Ausgabewert die Wahrscheinlichkeit einer bestimmten Klassenzugehörigkeit ausdrückt (IU Internationale Hochschule, 2025).
- **Naive Bayes**: Dieser Klassifikator beruht auf dem Satz von Bayes und geht von der Annahme aus, dass die Verteilungen der berücksichtigten Merkmale voneinander unabhängig sind (IU Internationale Hochschule, 2025).
- **Support Vector Machine (LinearSVC)**: Das Verfahren basiert auf der Einpassung einer Klassifizierungsgrenze (Hyperebene) in den hochdimensionalen Feature-Raum, wobei neue Beobachtungen je nach Position relativ zur Hyperebene klassifiziert werden (IU Internationale Hochschule, 2025).
- **Random Forest**: Diese Ensemble-Methode bündelt einzelne Entscheidungsbäume durch Bagging zu einem starken Schätzer, dessen Vorhersage durch Aggregation der Einzelvorhersagen ermittelt wird (IU Internationale Hochschule, 2025).
- **Gradient Boosting**: Im Gegensatz zu Random Forest erfolgt das Training der Entscheidungsbäume sequenziell, wobei jeder Schätzer die Fehler des vorherigen korrigiert (IU Internationale Hochschule, 2025).

Die TF-IDF-Vektorisierung wurde mit bis zu 20.000 Features und Uni- sowie Bigrammen konfiguriert. Zusätzlich wurde eine Textvorverarbeitung mit Kleinschreibung, Entfernung von HTML-Tags, URLs und Sonderzeichen sowie Lemmatisierung durchgeführt.

2.6.2 Transformer-basierte Modelle

Für den Vergleich mit modernen Deep-Learning-Ansätzen wurden zwei BERT-basierte Modelle ausgewählt:

DistilBERT Sanh et al., 2019 ist eine komprimierte Version des ursprünglichen BERT-Modells Devlin et al., 2018, die durch Knowledge Distillation erzeugt wurde. Mit 66 Millionen Parametern ist DistilBERT etwa 40% kleiner als BERT, behält jedoch 97% der Sprachverständnisfähigkeiten bei. Diese Eigenschaften machen DistilBERT besonders geeignet für Projekte mit begrenzten Hardwareressourcen.

Als zweiter Ansatz wurde ein **vortrainiertes BERT-Modell für Sentimentanalyse** (nlptown/bert-base-multilingual-uncased-sentiment) im Zero-Shot-Modus getestet. Dieses Modell wurde bereits auf Produktbewertungen trainiert und ermöglicht eine Klassifikation ohne zusätzliches Fine-Tuning auf dem eigenen Datensatz.

2.7 Training

2.7.1 Erstes Training der klassischen Modelle

Das initiale Training erfolgte mit drei klassischen Modellen: Logistic Regression, Multinomial Naive Bayes und LinearSVC. Das Modell mit der höchsten Validierungsgenauigkeit wurde automatisch ausgewählt und auf dem Testdatensatz evaluiert.

Die Evaluation umfasste Accuracy, Classification Report und Confusion Matrix. Das beste Modell wurde abschließend gespeichert.

Die initiale Evaluation zeigte eine deutliche Asymmetrie in der Klassifikationsleistung. Während 5-Sterne-Bewertungen mit einem Recall von 97% und einer Precision von 78% zuverlässig erkannt wurden, offenbarten die mittleren Klassen erhebliche Schwächen.

Die Confusion Matrix in Table 4 verdeutlicht das Kernproblem: Das Modell tendierte dazu, Rezensionen aller Klassen als 5-Sterne-Bewertungen zu klassifizieren. So wurden beispielsweise 163 der 1-Stern-Bewertungen fälschlicherweise als 5 Sterne eingestuft, bei 4-Stern-Bewertungen waren es sogar 381 von 528.

Tabelle 4: Confusion Matrix des ersten Trainings

	Vorhergesagt				
Tatsächlich	1	2	3	4	5
1 Stern	370	3	10	10	163
2 Sterne	82	24	23	19	82
3 Sterne	55	8	49	34	152
4 Sterne	25	2	15	105	381
5 Sterne	37	2	12	41	2796

2.7.2 Aufbau der Experiment-Pipeline

Die Erkenntnisse aus dem ersten Training zeigten, dass systematische Experimente mit verschiedenen Konfigurationen erforderlich sind. Um eine effiziente Durchführung und Vergleichbarkeit zu gewährleisten, wurde eine modulare Pipeline-Architektur implementiert, die aus drei aufeinander aufbauenden Komponenten besteht.

Die erste Komponente bildet das Preprocessing-Modul, das die Rohtexte für die maschinelle Verarbeitung vorbereitet. Folgende Verarbeitungsschritte können dabei flexibel aktiviert oder deaktiviert werden: Konvertierung zu Kleinbuchstaben, Entfernung von HTML-Tags, URLs, Zahlen und Satzzeichen sowie die Filterung englischer Stoppwörter. Für die Wortnormalisierung stehen zwei Alternativen zur Verfügung: der Porter Stemmer, der Wörter auf ihren Stamm reduziert, sowie der WordNet Lemmatizer, der Wörter auf ihre Grundform zurückführt. Zusätzlich kann eine minimale Token-Länge definiert werden, um sehr kurze Wortfragmente auszuschließen.

Die zweite Komponente umfasst die Feature-Extraktion, die den vorverarbeiteten Text in numerische Vektoren transformiert. Hierbei kann zwischen TF-IDF-Vektorisierung und einfacher Count-Vektorisierung (Bag-of-Words) gewählt werden. Die Anzahl der Features ist zwischen 5.000 und 20.000 konfigurierbar. Durch die Einstellung der N-Gramm-Range lassen sich ausschließlich Unigramme, zusätzlich Bigramme oder auch Trigramme berücksichtigen. Weitere Parameter wie minimale und maximale Dokumentfrequenz, sublineare TF-Skalierung sowie die IDF-Gewichtung ermöglichen eine Feinabstimmung der Feature-Repräsentation.

Die dritte Komponente stellt das Training-Modul dar, das fünf verschiedene Klassifikationsverfahren unterstützt: Logistic Regression, Multinomial Naive Bayes, Linear Support Vector Machine, Random Forest und Gradient Boosting. Für jedes Modell sind sinnvolle Standardparameter hinterlegt, die bei Bedarf überschrieben werden können.

Die gesamte Konfiguration erfolgt über YAML-Dateien, wodurch Experimente reproduzierbar dokumentiert werden. Jedes Experiment erhält automatisch einen Zeitstempel und speichert alle Ergebnisse strukturiert ab, einschließlich der verwendeten Konfiguration, aller Metriken, der Confusion Matrix sowie des trainierten Modells. Ein zusätzliches Vergleichsskript ermöglicht die tabellarische Gegenüberstellung aller durchgeführten Experimente.

3 Fazit

3.1 Zielerreichung und Projektergebnisse

3.2 Kritische Reflexion und gewonnene Erkenntnisse

3.3 Verbesserungspotenziale und Optimierungsansätze

3.4 Ausblick

Literatur

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2024). Bridging Language and Items for Retrieval and Recommendation. *arXiv preprint arXiv:2403.03952*.

IU Internationale Hochschule. (2025). *Maschinelles Lernen – Supervised Learning* [Kurscode: DLBDSMLSL01_D].

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schaaff, K. (2025). *Einführung in Natural Language Processing* [Kurscode: DLBAIINLP01_D]. IU Internationale Hochschule GmbH.

Verzeichnis der Anhänge