Projekt: NLP DLBAIPNLP01_D

Projektbericht

Studiengang: Angewandte Künstliche Intelligenz

Sven Behrens

Matrikelnummer: 42303511

Tutor: Prof. Dr. Maja Popovic

12. Dezember 2025

Inhaltsverzeichnis

Abbildungsverzeichnis **II**

Tabellenverzeichnis **III**

Abkürzungsverzeichnis **IV**

1 Einleitung **1**

2 Hauptteil **2**

2.1 Hardwareauswahl 2

2.2 Projektumgebung einrichten 2

2.3 Datenbeschaffung 2

2.4 Modellauswahl 3

2.4.1 Klassische Machine-Learning-Modelle 3

2.4.2 Transformer-basierte Modelle 4

2.5 Trainingsläufe 4

3 Fazit **4**

3.1 Zielerreichung und Projektergebnisse 4

3.2 Kritische Reflexion und gewonnene Erkenntnisse 4

3.3 Verbesserungspotenziale und Optimierungsansätze 4

3.4 Ausblick 4

Literaturverzeichnis **5**

Verzeichnis der Anhänge **6**

Anhang **6**

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 Gesamt-

datensatz von 150.000 Datensätzen entstand. Die Daten wurden stratifiziert nach Sternebewertung in Trainings- (70%), Validierungs- (15%) und Testdaten (15%) aufgeteilt.

2.4 Modellauswahl

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

2.4.1 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.

2.4.2 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.

2.5 Trainingsläufe

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*.

Verzeichnis der Anhänge

Anhang