Seminararbeit

M.Sc.-Studiengang "'Betriebswirtschaftslehre"'

Universität Hamburg
Fakultät für Betriebswirtschaft
Lehrstuhl für Statistik mit Anwendung in der Betriebswirtschaftslehre
Prof. Dr. Martin Spindler

Data Profiling mit Large Language Models - Eine Studie am Beispiel des US Gender-Pay-Gap

Gutachter: Prof. Dr. Martin Spindler

eingereicht von:
Vincent-Konstantin Kapp
7778027
Methfesselstraße 16
20257 Hamburg
0152-56761933

betreut von:
Prof. Dr. Martin Spindler

Abgabedatum, Ort:
07.01.2025, Hamburg

Einleitung

Abstract

Datengetriebene Prozesse können Unternehmen nicht nur dabei helfen, ihre Prozesse zu optimieren, sondern auch bei der Entwicklung neuer Strategien. Um festzustellen, welche Daten in den Datenbanken vorliegen und wie diese zueinander stehen, braucht es Mitarbeiter*innen, die einzeln überprüfen, welche Daten vorhanden sind und in welchen Zusammenhängen sie basierend auf Common Sense existieren. Data Profiling stellt Unternehmen mit historisch gewachsenen Datenbanksystemen vor Herausforderungen für die datengestützte Transformation. Können die aktuellen Entwicklungen von Large Language Modellen (LLMs) dazu beitragen, automatisiert die Zusammenhänge von Spalten zu erkennen? Um diese Frage zu beantworten, widmet sich diese Seminararbeit, aufbauend auf der Arbeit von Trummer (2024), der Untersuchung, wie gut LLMs anhand einer Case Study die Korrelation von Daten erkennen können. Dazu wird ein LLM auf Basis von GPT-3 trainiert und mit einem Benchmark-Verfahren verglichen. Die Ergebnisse zeigen, dass LLMs eine hohe Genauigkeit bei der Korrelationserkennung aufweisen und damit Unternehmen bei der Datenanalyse unterstützen können.

Table of contents

List of Figures

List of Tables

Chapter 1

Hintergrund

1.1 Data Profiling

1.2 Erkennen von Korrelationen in Datens

(Trummer 2023).

1.3 Language Models

(Arora et al. 2023).

1.4 Nutzung von NLP für Datenbankanalyse

1.5 Heterogenität in Datensätzen: am Beispiel des US Gender-Pay-Gap

Chapter 2

Methodik

2.1 Setup der Studie

2.1.1 Datenbasis und Vorbereitung

2.1.2 Vorgehensweise zur Korrelationsanalyse mit LLM´s

2.2 Benchmark-Metriken und Analysea

2.3 Benchmark-Analyse

Meine Überlegung ist es das Ergebnisse von (Bach, Chernozhukov, and Spindler 2024).

Chapter 3

Ergebnisse

3.1 Darstellung der Ergebnisse

3.2 Vergleich der Ergebnisse

Chapter 4

Diskussion

4.1 Diskussion der Ergebnisse

4.2 Stärken und Limitationen der Arbeit

4.3 Praktische Anwendungsmöglichkeit

Chapter 5

Eidesstattliche Erklärung

Eidesstattliche Erklärung

"Ich versichere, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und mich anderer als der im beigefügten Verzeichnis angegebenen Hilfmittel nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen übernommen wurden, sind als solche kenntlich gemacht. Alle Internetquellen sind der Arbeit beigefügt. Des Weiteren versichere ich, dass ich die Arbeit voher nicht in einem anderen Prüfungsverfahren eingereicht habe und dass die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht."
~\[20mm] Ort, Datum Unterschrift

References

Arora, Simran, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. 2023. "Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes." Journal Article. *arXiv Preprint arXiv:2304.09433*.

Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2024. "Heterogeneity in the US Gender Wage Gap." *Journal of the Royal Statistical Society Series A: Statistics in Society* 187 (1): 209–30.

Trummer, Immanuel. 2023. "Can Large Language Models Predict Data Correlations from Column Names?" Journal Article. *Proceedings of the VLDB Endowment* 16 (13): 4310–23.