Optimierung von Large Language Model-basierten Datenextraktionsprozessen: Ein systematischer Ansatz zur Klassifizierung interner E-Mails

SPERRVERMERK

Zweite Projektarbeit

aus dem Studiengang Wirtschaftsinformatik Sales & Consulting

an der Dualen Hochschule Baden-Württemberg Mannheim

von

Tim Christopher Eiser

Bearbeitungszeitraum:	Datum - Datum
Matrikelnummer, Kurs:	Matrikelnummer, Kurs
Studiengangleiter:	Name
Ausbildungsfirma:	SAP SE Dietmar-Hopp-Allee 16 69190 Walldorf, Deutschland
Betreuer der Ausbildungsfirma:	Name E-Mail Telefonnummer
Wissenschaftliche Betreuerin:	Name E-Mail Telefonnummer

I. Eidesstattliche Erklärung

Ich versichere hiermit, dass ich meine Projektarbeit mit dem Thema: „Titel“ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort, Datum	**Unterschrift**

II. Sperrvermerk

Die nachfolgende Arbeit enthält vertrauliche Daten und Informationen der SAP SE, Dietmar-Hopp-Allee 16, 69190 Walldorf, Deutschland. Der Inhalt dieser Arbeit darf weder als Ganzes noch in Auszügen Personen außerhalb des Prüfungsprozesses und des Evaluationsverfahrens zugänglich gemacht werden. Veröffentlichungen oder Vervielfältigungen der Projektarbeit - auch auszugsweise - sind ohne ausdrückliche Genehmigung der SAP SE in einem unbegrenzten Zeitrahmen nicht gestattet. Über den Inhalt dieser Arbeit ist Stillschweigen zu wahren.

SAP und die SAP Logos sind eingetragene Warenzeichen der SAP SE. Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in dieser Arbeit berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedem benutzt werden dürfen.

III. Gleichbehandlung der Geschlechter

In dieser Praxisarbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten werden dabei ausdrücklich mitgemeint, soweit es für die Aussage erforderlich ist.

IV. Disclaimer

Ein Teil der Literatur, die für die Anfertigung dieser Arbeit genutzt wird, ist nur über die E-Book-Plattform o'Reilly abrufbar. Bei diesen Ressourcen existieren keine Seitennummern, es wird bei Verweisen stattdessen die Kapitelnummer angegeben.

Um den Lesefluss zu verbessern, werden Abbildungen, Codebeispiele und Tabellen, die den Lesefluss stören, im Anhang platziert, auf den im Text zusätzlich verwiesen wird.

V. Abstract

Titel:	Optimierung von Large Language Model-basierten Datenextraktionsprozessen: Ein systematischer Ansatz zur Klassifizierung interner E-Mails
Verfasser:	Tim Christopher Eiser
Kurs:	WWI 23 SCB
Ausbildungsbetrieb:	SAP SE

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Codeverzeichnis

Abkürzungsverzeichnis

Test	Test

Variablenverzeichnis

Q	Test

1. Einleitung

1.1. Motivation

Die Einleitung erläutert die zunehmenden Herausforderungen bei der Bearbeitung großer Mengen unstrukturierter interner Kommunikation in Unternehmen. Ziel ist die Anwendung und Optimierung von LLMs (Large Language Models) zur Verbesserung der Effizienz und Genauigkeit der Datenextraktion und -klassifizierung in E-Mails.

1.2. Ziel und Gang

- Zielsetzung: Systematische Optimierung der Datenextraktion und -klassifizierung durch LLMs und gezieltes Prompt Engineering.
- Gang der Untersuchung: Die Arbeit basiert auf dem CRISP-DM-Zyklus, der die Schritte von der Problemdefinition über die Datenverarbeitung und Modellierung bis zur Evaluation und praktischen Umsetzung umfasst.

2. Methodik

Die Methodik beschreibt die Anwendung des CRISP-DM-Zyklus:

- Business Understanding (Geschäftsverständnis): Definition der Projektziele und des geschäftlichen Nutzens.
- Data Understanding (Datenverständnis): Analyse der verwendeten Datenquellen, insbesondere interner E-Mails.
- Data Preparation (Datenaufbereitung): Reinigung, Formatierung und Extraktion relevanter Merkmale.
- Modeling (Modellierung): Entwicklung und Anpassung von LLM-Modellen zur Datenklassifikation mit Hilfe von Prompt Engineering.
- Evaluation (Bewertung): Überprüfung und Bewertung der Modellergebnisse.
- Deployment (Einsatz): Integration des Systems in bestehende Kommunikationsprozesse.

3. Grundlagen

- LLM-Technologien und ihre Funktionsweise: Erklärung der Grundlagen von Large Language Models (z.B. GPT, BERT) und ihrer Einsatzmöglichkeiten.
- Prompt Engineering: Techniken und Methoden zur Leistungsoptimierung von LLMs durch gezielte Gestaltung von Eingabeaufforderungen.
- CRISP-DM-Zyklus: Beschreibung des CRISP-DM-Frameworks als strukturierter Ansatz für datengetriebene Projekte.

4. Praxis

- Datenerhebung und -aufbereitung: Sammlung und Bearbeitung von E-Mail-Daten unter Berücksichtigung des Datenschutzes, explorative Datenanalyse zur Identifikation relevanter Merkmale.
- Modellierung und Entwicklung: Implementierung und Anpassung von LLM-Modellen zur Klassifikation von E-Mails, Anwendung verschiedener Prompting-Strategien.
- Integration und Evaluierung: Einsatz des optimierten Modells und Analyse der Ergebnisse im Vergleich zu bestehenden Ansätzen.

5. Diskussion

- Ergebnisse und Interpretation: Analyse der Ergebnisse und Vergleich mit bestehenden Ansätzen zur Datenextraktion und -klassifizierung.
- Kritische Bewertung: Reflexion über die Anwendung des CRISP-DM-Zyklus und Identifikation von Stärken, Schwächen und Verbesserungsmöglichkeiten.

5.1. Einordnung der Ergebnisse

5.2. Herausforderungen und Limitationen

5.3. Handlungsempfehlungen und zukünftige Forschung

6. Schlussbetrachtung

- Zusammenfassung der Ergebnisse: Darstellung der erreichten Verbesserungen durch LLMs.
- Ausblick: Empfehlungen für zukünftige Forschung und Weiterentwicklung von LLM-basierten Systemen zur Optimierung interner Kommunikationsprozesse.

6.1. Fazit

6.2. Ausblick

i. Literaturverzeichnis

ii. Anhang

iii. Anhang A

iv. Anhang B

v. Anhang C