Künstliche Intelligenz im Kundenservice: Ein Vergleich generalistischer LLMs und spezialisierter Modelle zur Textklassifikation

Kerim Kaan Özkara
Institut für Informatik, Humboldt-Universität zu Berlin
Gutachter: Prof. Dr. Ulf Leser
Zweitgutachter: ausstehend

Einleitung

Die Grundlage für diese Bachelorthesis ist ein für die Vattenfall Europe Sales GmbH durchgeführtes Projekt. Im Rahmen des Projektes werden verschiedene Möglichkeiten untersucht, wie Kundentexte effizient klassifiziert werden können. Um sowohl den Datenschutz als auch realitätsnahe Evaluationsergebnisse zu gewährleisten, werden die untersuchten KI-Modelle in der Unternehmensumgebung mit realen Kundendaten trainiert, getestet und validiert. In der Thesis verwendete Datenbeispiele werden anonymisiert.

Im Folgenden werden Modelle mit mehr als einer Milliarde Parametern als große Sprachmodelle (Large Language Models, LLMs) bezeichnet. Sie werden aufgrund ihrer Größe und der benötigten Rechenleistung häufig über die Application Programming Interface (API) eines Modellbetreibers genutzt. Mit kleinen Modellen dagegen sind im Folgenden Modelle gemeint, die bis zu 1 Milliarde Parameter haben. Oft werden sie ohne GPU lokal oder auf kostengünstigen Rechenclustern in einer Cloud trainiert und betrieben.

Relevanz des Themas

Die steigende Anzahl von Kundenanfragen und die Bearbeitungskosten pro Anfrage (cost to serve = cts) stellen für Vattenfall eine wachsende Herausforderung dar. Das folgende Beispiel soll das verdeutlichen. Täglich erhält Vattenfall über das Kundenserviceportal durchschnittlich mehrere Tausend schriftliche Kundenanfragen. Die manuelle Bearbeitung durch einen Kundenservicemitarbeiter nimmt mehrere Minuten pro Anfrage in Anspruch. Dafür muss der Kundenservicemitarbeiter die konkreten Anliegen in den Kundenanfragen erkennen und die passenden Prozesse zur Bearbeitung auswählen. Nachdem die für die Prozessausführung nötigen Informationen aus dem Text oder dem Anhang extrahiert worden sind, werden diese ins bestehende System eingetragen.

Um die Servicequalität den Kunden gegenüber zu erhöhen, sucht Vattenfall nach effizienten Lösungen. Die Lösung soll Fehlinterpretationen minimieren, den korrekten Prozess einleiten und die Antwortzeiten reduzieren. All dies soll zudem zur Entlastung von Mitarbeitern sowie Führungskräften führen und Kosten senken.

Der erste Schritt zur Implementierung einer Lösung muss daher das Erkennen der konkreten Anliegen aus den Kundenanfragen sein. Daher gewinnen Verfahren der automatisierten Textverarbeitung, insbesondere Natural Language Processing (NLP), zunehmend an Bedeutung [1].

Sprachmodelle

Innerhalb von NLP kommen häufig LLMs wie GPT-4 zum Einsatz. Sie ermöglichen Textanfragen automatisiert zu kategorisieren, für die Prozesslogik relevante Informationen zu extrahieren und dem Kunden automatisch zu antworten [2]. Viele Unternehmen pilotieren bereits jetzt solche Modelle zur Unterstützung im Kundensupport [3].

Die in der Bachelorthesis verwendeten Personenbezeichnungen beziehen sich immer gleichermaßen auf weibliche und männliche Personen. Auf eine Doppelnennung und gegenderte Bezeichnungen wird zugunsten einer besseren Lesbarkeit verzichtet.

Der Einsatz von LLMs wie GPT-4 im Kundenservice kann die Produktivität und Servicequalität steigern [4, 5, 2]. Closed-Source-LLMs stellen Informationen, wie die Model Weights, Trainingsdaten und den zum Training verwendeten Quellcode oftmals nicht zur Verfügung [6]. Solche Closed-Source-Modelle wie GPT-4 sind zwar in die bestehende IT-Landschaft von Vattenfall einfacher zu implementieren, weisen gegenüber Open Source Lösungen folgende Nachteile auf. Sie haben hohe Betriebskosten, die direkt von der Anzahl der Kundenanfragen abhängen. Auch die Abhängigkeit vom Modellanbieter und datenschutzrechtliche Bedenken sind nicht zu unterschätzende Nachteile.

Eine Alternative stellen kleinere, auf unternehmensspezifischen Daten trainierte Modelle dar. Diese können aufgrund ihrer geringeren Hardwareanforderungen lokal oder auf vergleichsweise günstigen Rechenclustern in der Cloud trainiert werden, um ihre Leistung für spezifische Aufgaben zu maximieren (Fine-Tuning) [7]. Dadurch lassen sich die laufenden und initialen Kosten reduzieren. Zudem bieten sie eine höhere Flexibilität hinsichtlich Anpassungen und Optimierungen.

Forschungsstand

Textklassifikation

Die Textklassifikation ist ein Teilgebiet des Natural Language Processings (NLP). Sie bezeichnet die Zuordnung von Texten in vordefinierte Kategorien. Dabei kommen traditionelle Machine-Learning-Modelle und -Algorithmen wie Naive Bayes (NB), Support Vector Machines (SVM) und Random Forest (RF) sowie moderne Deep-Learning-Ansätze wie Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) und Transformer-Modelle wie BERT und RoBERTa zum Einsatz [8, 9, 10].

Zentrale Herausforderungen der Textklassifikation für Vattenfall sind die Verfügbarkeit und Qualität der Trainingsdaten sowie die Komplexität in der Ausdrucksweise und Anliegen der Kunden. Insbesondere die Multiklassenklassifikation sowie die Unterscheidung semantisch ähnlicher Klassen stellen eine große Herausforderung dar [11, 12]. Auch die technischen und finanziellen Limitationen bergen besondere Herausforderungen.

Durch das Fine-Tuning werden Large Language Models auf spezifische Klassifikationsaufgaben angepasst, was zu einer signifikanten Steigerung der Klassifikationsleistung führt [13, 14].

Vektorembeddings

Vektorembedding bezeichnet den Prozess der Umwandlung von Wörtern in Vektoren. Dabei wird die Semantik eines Wortes im Vektor abgebildet. „Word embeddings have demonstrated their effectiveness in storing valuable syntactic and semantic information"[15]. Sie können für verschiedene NLP-Aufgaben, darunter auch Textklassifikation, eingesetzt werden [16]. Es gibt verschiedene Verfahren, um Embeddings zu erzeugen. Kontextunabhängige Verfahren können nur die Bedeutung jedes einzelnen Wortes in einen Vektor umwandeln. Dadurch entsteht aber das Problem der Conflation Deficiency, die die Doppeldeutigkeit eines Wortes und die deshalb ungünstig umgewandelten Vektoren beschreibt [15]. Kontextabhängige Ansätze betrachten jedes Wort im Kontext und sorgen so für eine genauere Umwandlung der Bedeutung eines Wortes im Satz in einen Vektor. Zu kontextabhängigen Verfahren gehören zum Beispiel word2vec und GloVe [17]. BERT ist ein

Beispiel für einen transformerbasiertes kontextabhängiges Embeddingmodell [17].

Diese Vektoren können direkt als Input für ein Klassifikationsmodell genutzt oder persistent in einer Vektordatenbank gespeichert werden.

Large Language Models im Kundenservice

LLMs wie GPT-4 haben sich als vielseitige Werkzeuge in der NLP-Forschung und -Anwendung etabliert. Solche Modelle sind in der Lage, Textklassifikationsaufgaben durchzuführen, da sie auf einem großen Datensatz vortrainiert wurden und somit ein „profound understanding of natural language and the ability to generate coherent and contextually relevant responses"[18] haben. GPT-4 oder auch spezialisierte Reasoning-Modelle wie zum Beispiel GPT-o1, die über eine API eines nicht in der EU ansässigen Unternehmens erreichbar sind, verursachen hohe Betriebskosten, erfordern eine Anpassung an unternehmensspezifische Daten und werfen insbesondere für Unternehmen der kritischen Infrastruktur erhebliche Datenschutzbedenken auf [19]. Zudem führt der Einsatz von Closed-Source-Modellen oft zu einer erheblichen Abhängigkeit des Unternehmens vom Modellbetreiber. Dadurch ist die Anpassbarkeit der Lösung begrenzt und nur mit weiteren hohen Kosten umsetzbar.

In der Energiewirtschaft spielt die Datensicherheit eine besonders zentrale Rolle, und der Einsatz großer Modelle muss stets in Übereinstimmung mit regulatorischen Anforderungen und Datenschutzrichtlinien erfolgen. Aktuell steigen die Anforderungen durch neue Gesetze, wie dem EU AI Act und der daraus resultierenden KI-Verordnung [20].

Wegen der schwierigeren Skalierbarkeit und des höheren Wartungsaufwands will Vattenfall auf containerisierte Lösungen in der Cloud anstatt auf On-Premise-Lösungen setzen. Vattenfall setzt bereits eine Cloud-Plattform für verschiedene Anwendungen ein. Hier können weitere Rechencluster angemietet werden, um mit wenig Speicher eine Modellarchitektur aufzubauen. Die gewünschten Ausgaben sollen innerhalb weniger Sekunden erzeugt werden. Um die Kosten für die Cluster niedrig zu halten, soll die Modellausführung (Inference) auf einer CPU oder einer günstigen GPU betrieben werden.

Prompt Engineering

„A prompt is a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities."[21]

Prompt-Engineering bezeichnet das Anpassen der Eingabe für ein Sprachmodell, so dass die Antwort möglichst relevant und präzise ist [21, 22]. Die manuelle Erstellung geeigneter Prompts erfordert anfangs einen gewissen Aufwand. Der große Vorteil dabei ist, dass dadurch die großen LLMs für spezifische Anwendungsfälle optimiert werden können, ohne ein eigenes Training durchführen zu müssen.

Forschungsfrage und Zielsetzung

Die zentrale Forschungsfrage dieser Arbeit lautet: „Können kleinere, spezialisierte NLP-Modelle durch gezielte Optimierungen, Fine-Tuning und Kombinationen eine vergleichbare oder bessere Leistung bei der Textklassifikation als große LLMs wie GPT-4 erzielen?"

Die Evaluationskriterien lauten: Präzision, Recall, F1-Score, Implementierungsaufwand, Kosten und Datenschutz.

Ziel ist es, eine Entscheidungsgrundlage für den Einsatz geeigneter Sprachmodelle im Kundenservice kritischer Infrastrukturen zu schaffen.

Methodik

Datenbasis

Der Datensatz stammt aus dem Kundenserviceportal des Unternehmens. Jeder Datenpunkt enthält Header- und Bodyinformationen. Headerinformationen sind Anrede, Name, E-Mail-Adresse und bei eingeloggten Nutzern die Kontonummer. Die Bodyinformation ist der vom Kunden verfasste Text.

Das ist ein anonymisiertes Beispiel eines Kundentextes ohne Headerinformationen:

„Sehr geehrte Damen und Herren, mein Zählerstand für Vertragskonto 123 456 789 000 mit der Zählernummer 12 345 678 900 00 beträgt am 31.01.2024 012345. Mit freundlichen Grüßen Max Mustermann"

Die Kunden können innerhalb einer E-Mail ein oder mehrere Anliegen zum Ausdruck bringen. Da mehrere Anliegen innerhalb einer E-Mail oftmals komplexere Zusammenhänge haben, müssen diese teilweise oder vollständig von Kundenservicemitarbeitern bearbeitet werden. Daher sollen vorerst nur Modelle auf einer Ausgabeklasse trainiert werden. Insgesamt gibt es im Datensatz 61 mögliche Klassen, die in die folgenden Oberklassen eingeteilt sind:

1. Anfrage
2. Änderung
3. Feedback
4. Korrektur
5. Sonderthemen
6. Zahlung
7. Mitteilung
8. Beratung

Die Kundentexte werden regelmäßig von Mitarbeitern in eine oder mehrere Klassen eingeordnet. Diese Annotationen bilden beim Training die Ground Truth. Kundenanfragen, die in mehrere Kategorien eingeordnet wurden, werden nicht in den Datensatz aufgenommen.

Der verwendete Datensatz besteht aus 6626 Datenpunkten. Ein Teil des Datensatzes mit 2485 Datenpunkten wurde Im Dezember 2024 erstellt, ein weiterer Teil mit 2193 Datenpukten wurde im September 2024 erstellt und im März 2025 aufbereitet, der dritte Teil mit 1948 Datenpunkten wurde im März 2025 erstellt. Dies hat den Hintergrund, dass je nach Zeitraum die Art der Anfragen variieren kann und so ein zeitlicher Bias vermieden wird.

Modellentwicklung

Es werden drei Ansätze entwickelt und trainiert oder gepromptet: ein Logistic-Regression-Modell auf erzeugten Embedding-Vektoren, ein kleines, fine-tuned LLM (zum Beispiel BERT) sowie ein über eine externe API aufgerufenes und gepromptetes LLM (zum Beispiel GPT-4). Letzteres ist ein vom Unternehmen in der EU gemietetes und für die verwendeten Kundendaten freigegebenes Modell. Es wird auch die Möglichkeit untersucht Modelle wie BERT auf weniger Ausgabeklassen zu trainieren und hierarchisch hintereinanderzuschalten.

Evaluation

Die Metriken werden anschließend analysiert und die jeweilige Modelleignung für das Unternehmen daraufhin evaluiert. Die Evaluationskriterien umfassen Präzision, Recall, F1-Score, Implementierungsaufwand, laufende Kosten, Trainingszeit sowie die Skalierbarkeit und Optimierbarkeit der Modelle. Ein besonderer Fokus liegt auf der Untersuchung der Frage, inwiefern kleinere Modelle durch Kombination und Spezialisierung konkurrenzfähig zu einem großen LLM sein können.

Schließlich werden die Modelle hinsichtlich der genannten Kriterien miteinander verglichen, ihr Optimierungspotenzial analysiert und ihre Praktikabilität im Kundenservice evaluiert. Außerdem werden verschiedene Szenarien angegeben, in denen der Einsatz einer anderen Modellstrategie von Vorteil sein könnte. So kann je nach Bedarf eines Unternehmens eine geeignete Modellstrategie ausgewählt werden.

Literatur

[1] N. Patel and S. Trivedi, "Leveraging predictive modeling, machine learning personalization, nlp customer support, and ai chatbots to increase customer loyalty," *Empirical Quests for Management Essences*, vol. 3, no. 3, pp. 1–24, 2020.

[2] J. Wulf and J. Meierhofer, "Utilizing large language models for automating technical customer support," 2024.

[3] J. Deng and Y. Lin, "The benefits and challenges of chatgpt: An overview," *Frontiers in Computing and Intelligent Systems*, vol. 2, no. 2, pp. 81–83, 2022.

[4] M. A. A. Daqar and A. K. Smoudy, "The role of artificial intelligence on enhancing customer experience," *International Review of Management and Marketing*, vol. 9, no. 4, p. 22, 2019.

[5] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, *et al.*, "Chatie: Zero-shot information extraction via chatting with chatgpt," *arXiv preprint arXiv:2302.10205*, 2024.

[6] S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek, "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms," 2024.

[7] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.

[8] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022.

[9] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *International Journal of Research in Marketing*, vol. 36, no. 1, pp. 20–38, 2019.

[10] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[11] G. Forman, "A pitfall and solution in multi-class feature selection for text classification," in *Proceedings of the twenty-first international conference on Machine learning*, p. 38, 2004.

[12] Y. H. Li and A. K. Jain, "Classification of text documents," *The Computer Journal*, vol. 41, no. 8, pp. 537–546, 1998.

[13] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," in *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pp. 194–206, Springer, 2019.

[14] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[15] J. Camacho-Collados and M. T. Pilehvar, "From word to sense embeddings: A survey on vector representations of meaning," *Journal of Artificial Intelligence Research*, vol. 63, pp. 743–788, 2018.

[16] Y. Kim, "Convolutional neural networks for sentence classification," 2014.

[17] C. Wang, P. Nulty, and D. Lillis, "A comparative study on word embeddings in deep learning for text classification," in *Proceedings of the 4th international conference on natural language processing and information retrieval*, pp. 37–46, 2020.

[18] Y. Ge, W. Hua, K. Mei, J. Tan, S. Xu, Z. Li, Y. Zhang, *et al.*, "Openagi: When llm meets domain experts," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5539–5568, 2023.

[19] K. Kheiri and H. Karimi, "Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning," *arXiv preprint arXiv:2307.10234*, 2023.

[20] Europäische Union, "Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 über künstliche Intelligenz und zur Änderung bestimmter Rechtsakte der Union (AI-Act)," 2024.

[21] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[22] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *arXiv preprint arXiv:2310.14735*, 2023.