MASTERTHESIS

Automated Analysis of Unstructured Data from PDF Files using Artificial Intelligence

zur Erlangung des akademischen Grades

Master of Science

eingereicht im Fachbereich Mathematik, Naturwissenschaften und Datenverarbeitung
an der Technischen Hochschule Mittelhessen

Jonas Wölfer

Studiengang: Wirtschaftsinformatik M.Sc.

Eingereicht am: 14.02.2024

Referent: Prof. Dr. Harald Ritz
Korreferent: Prof. Dr. habil. Markus Siepermann

Abstract

In recent decades, digitalization has increased in almost all areas of life, which has led to enormous data growth and brought with it both new challenges and opportunities (Statista 2023). Unstructured data, which is often found in PDF files, is difficult to access for automated analyses, making it difficult to use effectively. Against this backdrop, the development of technologies for the automated analysis of unstructured data has become increasingly important. The use of artificial intelligence to recognize patterns and correlations in the data and extract relevant information has proven to be particularly promising (Zhao et al. 2023, 1). An important milestone in this field was the publication of the attention mechanism in 2017, which represented a revolutionary approach to contextual text recognition (Vaswani et al. 2017). Since then, the technology has continued to evolve and is now facing new challenges, as shown by Zhao et al. (2023).

This thesis addresses the question of how unstructured data in PDFs can be captured and efficiently processed using modern image recognition methods to give companies an advantage. A particular focus is placed on Natural Language Processing technologies, which serve as the basis for analyzing and processing the data. Finally, a prototype is used to show how the knowledge gained can be applied in practice.

Zusammenfassung

In den letzten Jahrzehnten hat die Digitalisierung in nahezu allen Lebensbereichen zugenommen, was zu einem enormen Datenwachstum geführt hat und sowohl neue Herausforderungen als auch Chancen mit sich bringt (Statista 2023). Insbesondere unstrukturierte Daten, wie sie häufig in PDF-Dateien vorliegen, sind für automatisierte Analysen schwer zugänglich, was ihre effektive Nutzung erschwert. Vor diesem Hintergrund hat die Entwicklung von Technologien zur automatisierten Analyse unstrukturierter Daten an Bedeutung gewonnen. Als besonders vielversprechend hat sich dabei der Einsatz von künstlicher Intelligenz erwiesen, um Muster und Zusammenhänge in den Daten zu erkennen und relevante Informationen zu extrahieren (Zhao et al. 2023, 1). Ein wichtiger Meilenstein in diesem Bereich war die Veröffentlichung des Aufmerksamkeitsmechanismus im Jahr 2017, der einen revolutionären Ansatz für die kontextbezogene Texterkennung darstellte (Vaswani et al. 2017). Seitdem hat sich die Technologie stetig weiterentwickelt und steht heute vor neuen Herausforderungen, wie Zhao et al. (2023) zeigen.

Die vorliegende Arbeit widmet sich der Frage, wie unstrukturierte Daten in PDFs mit modernen Methoden der Bilderkennung erfasst und effizient verarbeitet werden können, um Unternehmen einen Vorteil zu verschaffen. Ein besonderer Fokus liegt dabei auf Technologien des Natural Language Processing, die als Grundlage für die Analyse und Verarbeitung der Daten dienen. Abschließend wird anhand eines Prototyps gezeigt, wie eine praktische Anwendung der gewonnenen Erkenntnisse erfolgen kann.