Bachelorarbeit

...

von

Ricardo Valente de Matos

Matrikelnummer: 7203677
im Studiengang Wirtschaftsinformatik
der Fachhochschule Dortmund

Erstprüfer: Prof. Dr.-Ing. Guy Vollmer
Zweitprüfer: Stephan Schmeißer, M. Sc., Adessoplatz 1, 44269 Dortmund

Dortmund, den 9. Februar 2024

Motivation

Die Suche nach qualifizierten Mitarbeitern ist für Unternehmen von entscheidender Bedeutung, um wettbewerbsfähig zu bleiben und langfristigen Erfolg zu sichern. In einer Zeit, in der der Arbeitsmarkt zunehmend global und dynamisch wird, stehen Organisationen vor der Herausforderung, aus einer Vielzahl von Mitarbeitern diejenigen zu identifizieren, die am besten zu einem spezifischen Projekt im Unternehmen passen. Hier setzt die Entwicklung eines Recommender Systems zur Mitarbeiterempfehlung an. Ein solches System kann Unternehmen dabei unterstützen, den Prozess der Mitarbeiterrekrutierung und -auswahl zu optimieren. Durch die Berücksichtigung verschiedener Kriterien wie Qualifikationen, Fähigkeiten, Erfahrungen kann das Recommender-System dazu beitragen, die Auswahl effektiv zu filtern und diejenigen herauszufiltern, die am besten zu einem Projekt im Unternehmen passen. Ein solches System bietet außerdem den Vorteil, den Prozess der Mitarbeiterempfehlung zu automatisieren und zu beschleunigen. Dies ermöglicht Unternehmen, schneller auf offene Stellen zu reagieren und potenzielle Kandidaten zeitnah zu identifizieren. Dadurch wird die Effizienz der Mitarbeitersuche verbessert und die Qualität der Einstellungsentscheidungen erhöht.

Das Potenzial von Recommender Systems wurde auch bei *adesso* entdeckt und nun wird nach und nach Wege gesucht, KI-gestützte Systeme in die eigenen Prozesse zu integrieren. Im internen Projekt *adesso Staffing Advisor* wird an einem Recommender-System zur Mitarbeiterempfehlung für ausgewählte Projekte gearbeitet. Die Umsetzung der Recommender Systems bedient sich verschiedener KI-basierten Ansätze. Ein ganz entscheidender Schritt im Prozess der Mitarbeiterempfehlung ist die Vorverarbeitung der Bedarfsmeldungen. Diese sind eine wertvolle Informationsquelle, die Fachkräften helfen kann, die Empfehlungen effizienter zu gestalten, um dadurch wettbewerbsfähig zu bleiben. Allerdings sind diese oft umfangreich und komplex, was ihre effektive Nutzung erschwert.

Deshalb ist es entscheidend, effiziente Methoden und Techniken des Information Retrieval anzuwenden, um so relevante Informationen schnell und präzise aus Bedarfsmeldungen zu extrahieren. Die Extraktion wichtiger Schlüsselwörter, Phrasen und Themen ermöglicht es einen besseren Einblick in die Ziele, Methoden und Ergebnisse der Projekte zu bekommen. Dadurch können fundierte Entscheidungen bezüglich der Personalbesetzung getroffen und Ressourcen effizient genutzt werden.

Problemstellung

In einer immer stärker vernetzten und informationsreichen Welt stehen Organisationen vor der Herausforderung, relevante Informationen effizient aus umfangreichen Bedarfsmeldungen zu extrahieren. Obwohl diese Beschreibungen wichtige Einblicke in Ziele, Methoden und Ergebnisse liefern, können sie aufgrund ihres Umfangs und ihrer Komplexität schwer durchsuchbar und analysierbar sein. Die manuelle Identifizierung und Extraktion relevanter Inhalte ist zeitaufwendig und fehleranfällig. Daher stellt sich die Problemstellung:

Wie können wir effektive Methoden und Techniken des Information Retrieval und Data-Mining nutzen, um automatisiert relevante Inhalte aus Bedarfsmeldungen im spezifischen Software Entwicklungs-Kontext zu extrahieren und somit die Effizienz, Genauigkeit und Geschwindigkeit der Informationsgewinnung für Führungskräfte zu verbessern.

Ziele und Ergebnisse der Arbeit

Diese Ausarbeitung präsentiert eine umfassende Untersuchung zur Entwicklung eines automatisierten Systems zur Extraktion relevanter Inhalte aus Bedarfsmeldungen im Software-Entwicklungs-Kontext.

- Die erste Phase dieser Ausarbeitung besteht darin, eine klare Erwartungshaltung hinsichtlich der Anforderungen und Bedürfnisse der Stakeholder zu entwickeln. Hierfür werden Interviews mit Führungskräften durchgeführt, um die Erwartungen bezüglich einer „perfekten“ Bedarfsmeldung herauszuarbeiten. Diese dient als Grundlage für die weiteren Entwicklungs- und Evaluierungsphasen.

- Im Anschluss erfolgt eine eingehende Analyse der Techniken <was für Techniken> des Information Retrieval und Data-Mining, um die besten Ansätze zur Extraktion relevanter Inhalte zu identifizieren. Diese Analyse bildet die Grundlage für die Konzeptionierung einer Vorverarbeitung, das eine Kombination der erforschten Ergebnisse darstellt. Die Implementierung dieses Modells erfolgt durch den Aufbau einer Pipeline in Python, die eine effiziente Verarbeitung und Extraktion der Bedarfsmeldungen ermöglicht.

- Zur Evaluierung der Leistungsfähigkeit des entwickelten Systems werden reale Bedarfsmeldungen und Mitarbeiterinformationen verwendet. Dabei wird überprüft, inwiefern das Ergebnis der definierten Erwartungshaltung entspricht. Mithilfe von den Metriken *Precision*, *Recall* und *F1-Score* werden Abweichungen, Ähnlichkeiten und Anpassungen in Parametern analysiert, um Erkenntnisse darüber zu gewinnen, wie das System inhaltlich abschneidet und verbessert werden kann.

- (Schließlich wird eine vergleichende Untersuchung mit einem auf Large Language Model basierenden Vorverarbeitungsansatz durchgeführt. Dabei werden die Performance, Zeit und Ergebnisqualität des entwickelten Systems mit diesem alternativen Ansatz verglichen. Dieser Vergleich dient dazu, die Stärken und Schwächen des entwickelten Systems zu identifizieren und gegebenenfalls weitere Verbesserungen vorzunehmen.)

Vorgehen und Zeitplan

Ziel ist es die Arbeit im Mai fertig zu stellen. Die einzelnen Monatsziele können aus der nachfolgenden Tabelle entnommen werden.

Februar	• Durchführung der Interviews mit Fachkräften • Zusammentragung aller relevanter Information Retrieval- und Preprocessing-Ansätze
März	• Durchführung der Interviews mit Fachkräften • Formulierung der Anforderungen für Bedarfsmeldungen
April	• Entwicklung des Eigenen Preprocessing-Modells • Evaluierung der Ergebnisse
Mai	• Schluss schreiben • Korrekturen

ToDo: Aufbau der Arbeit anpassen <–hier

Aufbau der Arbeit

[8]

[20]

[4]

[6]

[3]

information filtering [13]

preprocessing [1]

——- spam-filter [19] [7] [22] —- TF-IDF (Term Frequency-Inverse Document Frequency): TF-IDF ist eine statistische Methode, die verwendet wird, um die Relevanz eines Begriffs in einem Dokument relativ zu einem Korpus von Dokumenten zu bestimmen. Wörter mit höheren TF-IDF-Werten werden als potenzielle Schlüsselwörter betrachtet. [2] [18]

Text-Ranking-Algorithmen: Text-Ranking-Algorithmen wie TextRank oder YAKE (Yet Another Keyword Extractor) verwenden Graphen-basierte Methoden, um Schlüsselwörter in einem Text zu identifizieren. Diese Algorithmen bewerten die Wichtigkeit von Wörtern basierend auf ihrer Verbindung zu anderen Wörtern im Text und extrahieren Schlüsselwörter entsprechend ihrer Rangfolge. [14] [23] [16]

N-Gramm-Analyse: N-Gramme sind Sequenzen von N aufeinanderfolgenden Wörtern in einem Text. Durch die Analyse von N-Grammen können häufig auftretende Phrasen oder Begriffe identifiziert werden, die potenzielle Schlüsselwörter darstellen. [17]

Part-of-Speech (POS) Tagging: POS-Tagging wird verwendet, um die grammatischen Kategorien von Wörtern in einem Text zu bestimmen. Durch die Berücksichtigung von Wörtern mit bestimmten POS-Tags wie Substantiven oder Adjektiven können relevante Schlüsselwörter extrahiert werden. [12] [15]

Regelbasierte Ansätze: Regelbasierte Ansätze verwenden vordefinierte Regeln oder Muster, um Schlüsselwörter zu identifizieren. Dies kann beispielsweise das Extrahieren von Wörtern sein, die häufig im Text vorkommen oder bestimmten Mustern entsprechen. -katalogisierung

Hybride Ansätze: Hybride Ansätze kombinieren verschiedene Methoden und Techniken, um eine genauere Extraktion von Schlüsselwörtern zu ermöglichen. Zum Beispiel könnte eine Kombination aus TF-IDF-Gewichtung und Text-Ranking-Algorithmen verwendet werden, um eine robuste Schlüsselwortextraktion zu erreichen.

1 Einleitung

Die allgemeine Bewusstheit von KI-gestützten Systemen ist in den letzten Jahren vor allem durch die Verbreitung von Large Language Models wie Chat-GPT gestiegen [5]. Sie können komplexe Fragen beantworten und kurze Texte zu Themen verfassen [5].

Das Thema Künstliche Intelligenz wird auch in der Europäischen Kommission diskutiert. Der AI Act behandelt Aspekte der Sicherheit und des Vertrauens bei der Nutzung von KI-Systemen [9].
„Um einen Beitrag zum Aufbau eines widerstandsfähigen Europas für die digitale Dekade zu leisten, sollten Menschen und Unternehmen in der Lage sein, die Vorteile von KI zu nutzen und sich gleichzeitig sicher und geschützt zu fühlen.“[11]
Das Gesetz behandelt einen Ansatz für vertrauenswürdige KI. Es werden Anforderungen gestellt, die neben Aspekten wie der Risikominderung und Qualität der Daten auch eine hohe Robustheit und Genauigkeit sicherstellen sollen [10].

Das Potenzial von KI wurde auch bei der Einbindung in Unternehmen entdeckt. Eine KI kann Arbeiten übernehmen, die für Beschäftigte eine Entlastung bedeuten können [21]. *adesso* hat auch das Potenzial erkannt und sucht nun nach und nach Wege, KI-gestützte Systeme in die eigenen Prozesse zu integrieren. Im internen Projekt *adesso Staffing Advisor* wird an einem Recommender-System zur Mitarbeiterempfehlung für ausgewählte Projekte gearbeitet. Die Umsetzung der Recommender Systems bedient sich verschiedener KI-basierten Ansätze. Da der Staffing-Prozess geschäftskritisch ist, ist es für adesso wichtig, die Qualität der Ergebnisse zu überprüfen und deinen geeignetsten Ansatz zu ermitteln.

1.1 Problemstellung

Recommender-Systeme existieren bereits seit geraumer Zeit. Es wurden viele Methoden und Metriken entwickelt, die eine Auskunft über die Qualität der Ergebnisse solcher Systeme liefern. Da das Recommender-System auf das spezifische Szenario einer Mitarbeiterempfehlung abzielt, ist es notwendig, die Ergebnisse angepasst auf den Staffing-Kontext zu evaluieren. Das KI-basierte Recommender-System des adesso Staffing Advisors verfolgt mehrere ähnlichkeitsbasierte Ansätze, von denen jeder seine Stärken und Schwächen hat. Die Zielgruppe des Systems sind die Fachkräfte. Diese haben eine Erwartungshaltung, das vom System erfüllt werden soll.

(Das Recommender System des adesso Staffing Advisors ist nicht transparenz)
(eventuell eine darstellung des systems?)

1.2 Ziele und Ergebnisse der Arbeit

Die vorliegende Arbeit hat das Ziel zu untersuchen, ob die Ergebnisqualität der Recommender-System-Ansätze des adesso Staffing Advisors für das Unternehmen adesso geeignet sind. Hierfür wird ein Konzept entwickelt, das schrittweise durch konkrete Methoden und Metriken die Genauigkeit und Ähnlichkeit im Staffing-Kontext erfasst und die Ergebnisqualität evaluiert. Das Ziel besteht darin, am Ende einen Ansatz auszuwählen, der im Staffing-Kontext einsetzbar ist und den vorher definierten Anforderungen entspricht.

- Dazu werden Methoden und Metriken zur Genauigkeit und Ähnlichkeit der Ansätze angewendet, wie beispielsweise die Kosinus-Ähnlichkeit, Pearson-Korrelation und Jaccard-Ähnlichkeit auf konkrete Testdaten
- Außerdem wird die Repräsentationsfähigkeit der Inputdaten zur Überprüfung der Qualität und Relevanz analysiert
- Das System soll praktisch anwendbar und vielseitig sein und einer vorher definierten Bewertung standhalten. Es soll auch Auskunft über die Sensitivität des Systems durch Änderungen in der Menge der Testdaten geben
- Schließlich sollen Fachkräfte, die den Staffing-Prozess manuell durchführen, Bewertungen und Feedback liefern, um potenziell notwendige Anpassungen an das System zu rechtfertigen

2 Grundlagen

2.1 Künstliche Intelligenz

1. Starke KI beinhaltet Problemlösungen genereller Art. Das, was am Ehesten an sowas heran kommt ist ChatGPT. Dennoch ist das Konzept einer starken KI ein Produkt aus Science-Fiction. Die Idee ist, dass die Maschine eine Art Bewusstsein hat und ein selbstständiges Verständnis unterschiedlicher Wissensbereiche entwickelt. 2. Schwache KI beinhaltet meist die Problemlösung einer konkreten Art. KI ist ein Konstrukt aus komplexen Algorithmen. Wenn von KI gesprochen wird, ist immer eine schwache KI gemeint.

2.2 Recommender Systems

2.3 Warum Testen und Überwachen der KI

1. KI-Systeme übernehmen bereits kritische Aufgaben. Identifizierung von Unfällen, Feuer oder Naturkatastrophen sind Aufgaben, die von einer KI schneller, besser und effizienter erledigt werden kann. Bei kritischen Prozessen ist es wichtig, dass die KI die vorgesehenden Leistungen erbringt.

3 Verwandte Arbeiten

4 Adesso Staffing Advisor

4.1 Aufbau des Projekts

4.2 Preprocessing

4.2.1 Keyword-Extraction

4.2.2 Normalizing

4.2.3 Large Language Models

4.3 KI-Modelle

4.3.1 spacy

4.3.2 sbert

4.4 Nutzung von Daten

-Welche Informationen der Ergebnisse des KI-Ansatzes sind vorhanden und werden gebraucht. (preprocessing, Ergebnis, similarity-Werte)

4.4.1 Welche Daten werden in die KI gegeben

4.4.2 Welche Daten werden von den KI-Ansätzen erstellt

4.4.3 Welche Daten werden von der KI zurückgegeben

5 Ähnlichkeitsmetriken

Ähnlichkeitsmetriken:
Überprüfe die Genauigkeit der Ähnlichkeitsmetriken, die im Recommender-System verwendet werden. Dazu gehören beispielsweise Kosinus-Ähnlichkeit, Pearson-Korrelation, Jaccard-Ähnlichkeit oder andere, je nach Kontext.
Top-N-Empfehlungen: Evaluieren Sie, wie gut das Recommender-System in der Lage ist, relevante Elemente unter den Top-N-Empfehlungen zu platzieren. Dies ist eine gängige Metrik, um die praktische Anwendbarkeit des Systems zu bewerten.
——————— Repräsentation der Merkmale: Untersuche, wie gut die Merkmale (Features) der Elemente im System repräsentiert sind. Eine gute Ähnlichkeitsberechnung hängt oft von der Qualität und Relevanz der Merkmale ab.
Diversität der Empfehlungen:
Prüfe, ob die Ähnlichkeitsbasierten Empfehlungen zu vielfältig sind. Eine zu starke Konzentration auf ähnliche Elemente könnte zu eintönigen Empfehlungen führen. Benutzerbewertungen und Feedback:
Integriere Benutzerbewertungen und -feedback in die Evaluierung, um sicherzustellen, dass die Ähnlichkeitsberechnungen den tatsächlichen Vorlieben der Benutzer entsprechen. Cold Start-Szenarien:
Teste das System unter Bedingungen des "Cold Start", um sicherzustellen, dass es auch effektive Empfehlungen machen kann, wenn es nur begrenzte Daten gibt. Auswirkungen von Merkmalen:
Analysiere, wie sich das Hinzufügen oder Entfernen von Merkmalen auf die Empfehlungen auswirkt. Dies kann helfen, die Sensitivität des Systems gegenüber verschiedenen Merkmalen zu verstehen. Nutzerinteraktion:
Es ist wichtig, die spezifischen Anforderungen deines Recommender-Systems zu berücksichtigen und die Evaluierungsmethoden entsprechend anzupassen. Kombiniere mehrere Metriken, um ein umfassenderes Bild der Leistung des Systems zu erhalten.

5.1 Genauigkeit der Ähnlichkeit

5.2 Qualität und Relevanz der Merkmale

5.3 Eintönige Empfehlungen

5.4 Benutzerbewertungen und -feedback

5.5 Cold Start

5.6 Sensitivität des Systems

6 Evaluation

-Evaluation der Art und Weisen der KI-Test- und Überwachungsmethoden (Wie hilfreich sind die unterschiedlichen Methoden, vielleicht mit einem Bewertungssystem im „adstaff lab“) Dashboard mit vielen verschiedenen Methoden der Visualisierung. Jede Methode hat einen eigenen Bereich, wo z.B. Sterne vergeben werden können.

7 Zusammenfassung und Ausblick

Ausblick

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt sowie die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Dortmund, den 9. Februar 2024

_____

Ricardo Valente de Matos

Literatur

[1] S. A. Alasadi und W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, Jg. 12, Nr. 16, S. 4102–4107, 2017.

[2] P. Bafna, D. Pramod und A. Vaidya, "Document clustering: TF-IDF approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, 2016, S. 61–66.

[3] N. J. Belkin und W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?" *Communications of the ACM*, Jg. 35, Nr. 12, S. 29–38, 1992.

[4] W. B. Croft, "Combining approaches to information retrieval," in *Advances in Information Retrieval: Recent Research from the center for intelligent information retrieval*, Springer, 2000, S. 1–36.

[5] L. De Angelis, F. Baglivo, G. Arzilli u. a., "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health," *Frontiers in Public Health*, Jg. 11, S. 1 166 120, 2023.

[6] R. Horesh, K. R. Varshney und J. Yi, "Information retrieval, fusion, completion, and clustering for employee expertise estimation," in *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, 2016, S. 1385–1393.

[7] A. Khorsi, "An overview of content-based spam filtering techniques," *Informatica*, Jg. 31, Nr. 3, 2007.

[8] M. Kobayashi und K. Takeda, "Information retrieval on the web," *ACM computing surveys (CSUR)*, Jg. 32, Nr. 2, S. 144–173, 2000.

[9] E. Kommission. "A european approach to artificial intelligence." (o. J.), Adresse: `https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence` (besucht am 3. Jan. 2024).

[10] E. Kommission. "Commission welcomes political agreement on artificial intelligence act*." (April 2021), Adresse: `https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473` (besucht am 3. Jan. 2024).

[11] E. Kommission. “Der Ansatz der EU für künstliche Intelligenz konzentriert sich auf Exzellenz und Vertrauen, um Forschung und industrielle Kapazitäten zu stärken und gleichzeitig Sicherheit und Grundrechte zu gewährleisten.” (o. J.), Adresse: `https://digital-strategy.ec.europa.eu/de/policies/european-approach-artificial-intelligence` (besucht am 3. Jan. 2024).

[12] D. Kumawat und V. Jain, “POS tagging approaches: A comparison,” *International Journal of Computer Applications*, Jg. 118, Nr. 6, 2015.

[13] C. Lanquillon, “Enhancing text classification to improve information filtering,” Diss., Otto-von-Guericke-Universität Magdeburg, Universitätsbibliothek, 2001.

[14] R. Mihalcea und P. Tarau, “Textrank: Bringing order into text,” in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, S. 404–411.

[15] T. Nakagawa und K. Uchimoto, “A hybrid approach to word segmentation and pos tagging,” in *Proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 2007, S. 217–220.

[16] T. Pay, S. Lucci und J. L. Cox, “An ensemble of automatic keyword extractors: TextRank, RAKE and TAKE,” *Computación y Sistemas*, Jg. 23, Nr. 3, S. 703–710, 2019.

[17] S. Pirk, “Implementierung und Visualisierung N-Gramm-basierter Word-Clouds,” B.S. thesis, 2019.

[18] J. Ramos u. a., “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, Citeseer, Bd. 242, 2003, S. 29–48.

[19] M. A. Shafi’I, M. S. Abd Latiff, H. Chiroma u. a., “A review on mobile SMS spam filtering techniques,” *IEEE Access*, Jg. 5, S. 15 650–15 666, 2017.

[20] A. Singhal u. a., “Modern information retrieval: A brief overview,” *IEEE Data Eng. Bull.*, Jg. 24, Nr. 4, S. 35–43, 2001.

[21] S. Stowasser, O. Suchy, N. Huchler u. a., “Einführung von KI-Systemen in Unternehmen,” *Gestaltungsansätze für das Change-Management. Whitepaper aus der Plattform Lernende Systeme, München*, 2020.

[22] K. Tretyakov, “Machine learning techniques in spam filtering,” in *Data Mining Problem-oriented Seminar, MTAT*, Citeseer, Bd. 3, 2004, S. 60–79.

[23] M. Zhang, X. Li, S. Yue und L. Yang, "An empirical study of TextRank for keyword extraction," *IEEE access*, Jg. 8, S. 178 849–178 858, 2020.