Thesis zur Erlangung des akademischen Grades Bachelor of Science (B. Sc.)

Automatisierung der Informationsgewinnung in Bedarfsmeldungen

von

Ricardo Valente de Matos

geboren am 30.10.1999
Matrikelnummer: 7203677
im Studiengang Wirtschaftsinformatik
der Fachhochschule Dortmund
im Fachbereich Informatik

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

Dortmund, den 24. April 2024

Abstract

ToDo: Abstract erstellen <–hie

Lesehinweis

Aus Gründen der besseren Lesbarkeit werden Wörter und Wortgruppen, die hervorgehoben werden oder mehrfach auftauchen, durch *kursiven* Text kenntlich gemacht. Zudem wird in dieser Ausarbeitung die Sprachform des generischen Maskulinums angewandt. Sämtliche Ausführungen sind jedoch geschlechtsunabhängig und beziehen sich damit auf alle Geschlechter.

Inhaltsverzeichnis

Abbildungsverzeichnis

Listings

1 Einleitung

In einer globalisierten und dynamischen Wirtschaftswelt sind Unternehmen zunehmend auf Projekte angewiesen, um ihre Ziele zu erreichen und Wettbewerbsvorteile zu erlangen. Die Personalbeschaffung für solche Projekte erfordert oft spezialisiertes Fachwissen und vielfältige Fähigkeiten, um erfolgreich umgesetzt zu werden. Es ist entscheidend für den Projekterfolg, dass die Personalbeschaffung die passenden Mitarbeiter für ausgewählte Projekte findet. 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 und 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 Führungskräften helfen kann, die Empfehlungen effizienter zu gestalten, um dadurch wettbewerbsfähig zu bleiben. Allerdings sind diese oft umfangreich, unsortiert 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.

1.1 Problemstellung

Um das Entlastungspotenzial für Führungskräfte durch das Gesamtsystem eines Recommender Systems für Mitarbeiterempfehlungen zu realisieren, sind mehrere Schritte notwendig. Eine Informationsgewinnung aus den unstrukturierten Projekt- und Mitarbeiterdaten ist unerlässlich, um schließlich den Ähnlichkeitsvergleich für die Empfehlungen durchführen zu können. Diese Ausarbeitung befasst sich mit dem ersten Schritt der Strukturierung und Informationsextraktion der vorhandenen *Bedarfsmeldungen*. Somit steht *adesso* 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.

In der Vergangenheit wurden bereits Methoden im Bereich des automatisierten Recruitings untersucht. Im Projektgeschäft sehen wir uns mit einem Problem konfrontiert, dessen Umfang jedoch präziser definiert werden kann, da die Kandidatenauswahl einem begrenzten Pool unterliegt. Besondere Relevanz hat hierbei die Erstellung einer Standardisierung der *Bedarfsmeldung*, da diese häufig unstrukturiert und mit fehlenden Informationen vorliegt.

1.2 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.

- In der Ausarbeitung wird zunächst ein Konzept einer standardisierten *Bedarfsmeldung* erarbeitet. Dazu wird eine klare Erwartungshaltung hinsichtlich der Anforderungen und Bedürfnisse der Stakeholder entwickeln. Hierfür werden Interviews mit Führungskräften durchgeführt, um die Erwartungen bezüglich

einer „perfekten“ *Bedarfsmeldung* herauszuarbeiten. Dieses Konzept dient als Grundlage für die weiteren Entwicklungs- und Evaluierungsphasen.

- Es wird an einer ausführbaren prototypischen Software gearbeitet, die *Bedarfsmeldungen* effizient verarbeitet und wichtige Informationen extrahiert. Hierfür wird eine Pipeline in Python aufgebaut und strukturell durch Use-Case- und UML-Diagramme dokumentiert. Es werden Modelle des Information Retrieval und Data-Mining implementiert. Dabei erfolgt zunächst eine eingehende Analyse der Techniken *TF-IDF*, *Text-Ranking-Algorithmen*, *N-Gramm-Analyse*, *POS-Tagging*, *Named Entity Recognition*, Regelbasierte Ansätze und Hybride Ansätze, um die besten Ansätze zur Extraktion relevanter Inhalte zu identifizieren. Diese Analyse bildet die Grundlage für die Konzeptionierung des Software-Prototypen, das eine Kombination der erforschten Ergebnisse darstellt.

- Um die Leistungsfähigkeit des entwickelten Systems zu evaluieren, werden Testfälle für reale *Bedarfsmeldungen* definiert. Dabei wird überprüft, inwieweit das Ergebnis den Erwartungen entspricht. Mit Hilfe einer manuellen Überprüfung werden Abweichungen, Ähnlichkeiten und Anpassungen analysiert, um Erkenntnisse über die inhaltliche Leistung des Systems und die Techniken zu gewinnen, die allein oder in Kombination mit mehreren Ansätzen die wichtigsten Informationen herausfiltern. Da die Dauer eine entscheidende Rolle spielt, werden auch Zeit und Leistung gemessen. Diese Ergebnisse werden mit einem neuen Vorverarbeitungsansatz verglichen, der auf dem Large Language Model basiert. Die Performance, Zeit und Ergebnisqualität des entwickelten Systems soll im Vergleich mit diesem alternativen Ansatz die Stärken und Schwächen des entwickelten Systems aufzeigen, um daraus gegebenenfalls weitere Verbesserungsmöglichkeiten zu identifizieren.

1.3 Aufbau der Arbeit

ToDo: Aufbau der Arbeit erstellen <–hie

g

2 Literaturüberblick

Das Ziel dieser Arbeit ist die Informationsgewinnung aus semistrukutrierten *Bedarfsmeldungen* für ein Recommender System, das Mitarbeiterempfehlungen innerhalb von *adesso* für ausgewählte Projekte generieren soll. In diesem Kapitel werden die für das Thema notwendigen Grundlagen und bereits erforschten Themengebiete im Kontext von Recommender Systemen und Informationsverarbeitung behandelt, die für das weitere Verständnis der Arbeit notwendig sind. Es wird ein Einblick in die Art und Weise gegeben, wie andere Autoren Information Retrieval und Filtering einsetzen und kombinieren.

2.1 Recommender Systems Historie und aktueller Stand der Forschung

Auch wenn die Erstellung eines Recommender Systems nicht Gegenstand der vorliegenden Ausarbeitung ist, stellt die Nutzung von Information Retrieval und Filtering ein entscheidener Schritt in Richtung eines funktionierenden Recommender Systems dar. Das Verständnis der Funktionsweise eines Recommender Systems sowie dessen Entwicklung in den vergangenen Jahren ist daher für das Verständnis des Teilbereichs dieser Thematik von Nutzen.

Recommender Systems existieren bereits seit vielen Jahren. Im Jahr 1992 führten Belkin und Croft eine Analyse und einen Vergleich des Information Retrievals und Filtering durch [10]. Das Information Retrieval behandelt dahingehend die grundlegende Technologie der Suchmaschine [10]. Das Recommender System basiert hauptsächlich auf der Technologie des Information Filtering. Im selben Jahr präsentierte Goldberg das Tapestry-System, welches das erste System zur Informationsfilterung darstellt, das auf kollaboratives Filtern durch menschliche Bewertung basiert. Die Mehrheit der frühen Empfehlungsmodelle basiert auf kollaborativer Empfehlungen, wobei K-Nearest-Neighbor (KNN)-Modelle eine besondere Rolle einnehmen. Diese Modelle prognostizieren die Nachbarn eines Zielnutzers, indem sie eine Ähnlichkeit zwischen den vorherigen Präferenzen und den Präferenzen der anderen Nutzer berechnen [10]. Die Studie von Goldberg inspirierte einige Forscher des Massachusetts Institute of Technology (MIT) und der University of Minnesota (UMN) dazu, einen Nachrichtenempfehlungsdienst mit dem Namen *GroupLens* zu entwickeln.

Die Hauptkomponente dieses Dienstes ist ein Modell zur kollaborativen Filterung zwischen Nutzern [10]. Das gleichnamige Forschungslabor kann somit als Pionier auf dem Gebiet der Recommender Systems bezeichnet werden. Die dort durchgeführten Forschungen bilden die Grundlage für nachfolgende Musik- und Video-Ähnlichkeitsempfehlungen [10].

Recommender Systeme haben in den letzten Jahren verschiedene Definitionen erhalten. Eine dieser Definitionen wird in dem Artikel von Resnick und Varian (1997) sinngemäß so beschrieben, dass ein typisches Recommender System Empfehlungen durch Personen als Eingabe erhält, die das System dann zusammenschließt und an geeignete Empfänger weiterleitet [5]. In einigen Fällen besteht die primäre Transformation in der Zusammenführung, in anderen Fällen liegt die Fähigkeit des Systems darin, gute Übereinstimmungen zwischen Empfehlungsgebern und Empfehlungsempfängern herzustellen [5]. Empfehlungssysteme stellen ein Instrument zur Interaktion mit umfangreichen und vielschichtigen Informationen dar. Sie ermöglichen eine personalisierte Sicht auf diese Informationen, indem sie die für den Nutzer wahrscheinlich relevanten Inhalte aufbereiten [5]. Besonders im Handelsverkehr im Internet sind Recommender Systeme ein häufiger Einsatzgebiet. Dabei werden Recommender Systeme als Werkzeuge zum Suchen und Filtern von Informationen verwendet, die dem Benutzer Vorschläge unterbreiten, die für ihn nützlich sein könnten. Sie sind in einer Vielzahl von Internetanwendungen weit verbreitet und helfen den Nutzern, bessere Entscheidungen bei der Suche nach Nachrichten, Musik, Urlaubsangeboten oder Geldanlagen zu treffen [33]. Eine spezifisches Recommender System konzentriert sich normalerweise auf eine Art von Themengebiet wie z. B. Filme oder Nachrichten [33]. Darüber hinaus sind sie zu einem entscheidenden Faktor in der Entscheidungsfindung von Organisationen geworden [6]. Unternehmen wie *adesso* bauen immer weiter auf Recommender System unterstützte System auf, um Prozesse zu beschleunigen oder zu vereinfachen.

Grundsätzlich können die Methoden in vier Typen unterteilt werden:

- collaborative Filtering-based (kollaborative Empfehlungssysteme)
- content-based (inhaltsbasierte Empfehlungssysteme)
- knowledge-based (wissensbasiert Empfehlungssysteme)
- hybrid (hybride Empfehlungssysteme)

Jede Empfehlungsmethode hat ihre Vorteile und Grenzen [22]. Insbesondere das inhaltsbasierte Empfehlungssystem bring eine hohe Relevanz für das Mitarbeiterempfehlungssystem. Die Grundprinzipien inhaltsbasierter Empfehlungssysteme sind zum einen die Analyse der Beschreibung der von einem bestimmten Benutzer bevorzugten *Items*, um die gemeinsamen Hauptattribute (Präferenzen) zu identifizieren, die diese *Items* unterscheiden. Diese Präferenzen werden in einem *Benutzerprofil* gespeichert [22]. Zusätzlich werden die Eigenschaften jedes *Items* mit dem *Benutzerprofil* verglichen, so dass nur *Items* empfohlen werden, die eine hohe Ähnlichkeit mit dem *Benutzerprofil* aufweisen [22]. Bei der Idee der Mitarbeiterempfehlung kann also die *Bedarfsmeldung* mit den benötigten Projektskills und Anforderung als *Benutzerprofil* angesehen werden. Die Mitarbeiterprofile sind dabei die *Items*. Die Attribute werden verglichen (Skills der Mitarbeiter mit den Skills und Anforderungen der *Bedarfsmeldung*) und ähnliche *Items* werden vorgeschlagen. Mit Hilfe traditioneller Methoden des Information Retrievals, wie z.B. dem Kosinus-Ähnlichkeitsmaß, werden dann Empfehlungen generiert [22]. Darüber hinaus generieren sie Empfehlungen mit Hilfe von statistischen und maschinelle Lernverfahren, die in der Lage sind, Nutzerinteressen aus historischen Nutzerdaten zu lernen [22].

2.2 Verwandte Arbeiten

Es gibt eine Reihe an verwandten Arbeiten die sich mit unterschiedlichen Aspekten des Staffing Prozesses und der Nutzung von Information Retrieval und Filtering zur Informationsgewinnung beschäftigen. Dennoch beschäftigt sich keine Arbeit mit dem spezifischen Problem der Informationsgewinnung aus *Bedarfsmeldungen*.

2.2.1 kp wie ich es nenne

Im ersten Paper beschreiben die Autoren einen Ansatz zur Ableitung von Unternehmensdaten und digitalen Fußabdrücken von Mitarbeitern. Mit Hilfe eines Big-Data-Workflows, der die Komponenten Information Retrieval und Suche, Datenfusion, Matrixvervollständigung und ordinale Regression nutzt, können Informationen zur Expertise automatisch zusammengeführt und für die Nutzung durch Experten aufbereitet werden. Das System soll Fähigkeiten, Talente und Fachwissens der Mitarbeiter in einem breiten Bereich wie cloud computing oder cybersecurity einschätzen. Beim Ansatz des Information Retrieval und -fusion wird eine Liste von Suchbegriffen erstellt, die sich auf das breite Fachgebiet der Mitarbeiter beziehen. Die Suche wird nach jedem dieser Abfragebegriffe durchgeführt, um Zusammenhänge zwischen Mitarbeiter und Datenquellen zu finden. Die verschiedenen Zusammenhänge werden

miteinander verschmolzen, gewichtet und nach der Abfrage sortiert. Die Mitarbeiter werden nach Daten gewichtet und bewertet, um einen einzigen Wert (sehr niedrig, niedrig, moderat, etwas, begrenzt) für ihr Fachwissen in diesem breiten Bereich zu erhalten.[14]

2.2.2 Information Filtering

Diese Arbeit befasst sich unter anderem mit dem Aspekt des content based Information Filtering. Das Ziel dabei ist es Informationen auf die Interessengebiete der Benutzer zu reduzieren. Dazu werden nicht relevante Dokumente aus einem Strom von Informationen entfernt, sodass dem Anwendern nur relevante Dokumente präsentiert werden. Ein Teil der Arbeit beschäftigt sich mit der Informationsfilterung und mögliche Filterungsvarianten werden vorgestellt. Die Arbeit konzentriert sich auf die inhaltsbasierte Filterung von Textdokumenten und identifiziert Informationsfilterung als einen Spezialfall der Textklassifikation. Dazu wird ein Überblick über gängige Methoden des Information Filtering gegeben und ihre Leistung evaluiert. [20]

2.2.3 Vorverarbeitung

Diese Arbeit zeigt Wege und Schritte zur Aufbereitung von Datensätzen auf. Die Arbeit umfasst Data-Mining Vorverarbeitungsmethoden, um die Qualität der Daten zu verbessern. Diese weisen wichtiger Schritte auf, um die Effizienz in der Datensammlung zu verbessern [1]. (Nicht sicher ob ich das drin lassen soll)

In diesem Beitrag wird der Teil des Anforderungsspezifikationsprozesses diskutiert, der zwischen der textuellen Anforderungsdefinition und den dazugehörigen Diagrammen der Anforderungsspezifikation liegt. Es wird die These aufgestellt, dass die Erstellung einer textuellen Anforderungsbeschreibung, welche das Verständnis des Analysten für das Problem darstellt, die Effizienz der Anforderungsvalidierung durch den Benutzer verbessert. Die vorliegende Idee ist aus dem Problem entstanden, dass Software-Entwickler nicht immer über die erforderlichen Kenntnisse in den fachlichen Abläufen der Themengebiete verfügen, die für die Erstellung der Software relevant sind. Im Rahmen der Anforderungsdefinition erfolgt eine textuelle Verfeinerung, welche als Anforderungsbeschreibung bezeichnet werden kann. Bei der Arbeit mit dem unterstützten Werkzeug *Tessi* ist der Analytiker durch die genannten Vorgaben gezwungen, Anforderungen zu vervollständigen und zu erklären sowie die

Rollen der Wörter im Text im Sinne der objektorientierten Analyse zu spezifizieren. Im Rahmen der Vorverarbeitung erfolgt eine Transformation der Requirements durch Templates.[18]

2.2.4 Hybride Ansätze

In der vorliegenden Untersuchung wird die Entwicklung von Kombinationen im Bereich des Information Retrievals analysiert. Dabei werden sowohl experimentelle Ergebnisse als auch die Retrieval-Modelle, die als formale Rahmen für die Kombination vorgeschlagen wurden, berücksichtigt. Es wird aufgezeigt, dass Kombinationsansätze für die Informationssuche als Kombination der Ergebnisse mehrerer Klassifikatoren auf der Grundlage einer oder mehrerer Darstellungen modelliert werden können. Zudem wird dargelegt, dass dieses einfache Modell Erklärungen für viele der experimentellen Ergebnisse liefern kann.[8]

Die vorliegende Arbeit kombiniert drei Ansätze des Information Retrievals mit dem Ziel, relevante Informationen aus Produktreviews zu extrahieren. Der Ansatz TF-IDF wird mit einem sogenannten CLASSIFIER Model kombiniert. Das Klassifikationsmodell verarbeitet drei Eingaben der Modelle LSTM, VADER und TF-IDF. Die Werte dieser Eingaben liegen im Bereich von [0,1]. Die Ausgabe des Klassifikationsmodells ist binär und gibt eine Vorhersage des vollständigen Textes der Modelleingabe aus (positiv oder negativ).[7]

Diese Arbeit befasst sich mit der Filterung von Fake news. In diesem Beitrag werden hybride Verfahren zur Gewinnung von Merkmalen untersucht, die in dem Gebiet noch nicht gründlich erforscht wurden. Die Anwendung von Hybridsystemen hat sich in einer Vielzahl von Anwendungsbereichen als nützlich erwiesen und zeigen eine Tendenz, die Fehlerquote zu reduzieren, indem sie Techniken wie TF-IDF und N-Grams verwenden.[35]

Im Rahmen dieser Studie wurde ein hybrider Algorithmus zur Extraktion von Schlüsselwörtern und Kosinusähnlichkeit zur Verbesserung der Satzkohäsion bei der Textzusammenfassung vorgeschlagen. Die vorgeschlagene Methode basiert auf einer Komprimierung von 50 %, 30 % und 20 %, um Kandidaten für die Zusammenfassung zu erstellen. Die Auswertung des Ergebnisses mittels t-Test zeigt, dass die vorgeschlagene Methode den Kohäsionsgrad signifikant erhöht. Der Ablauf umfasst die Analyse eines Dokuments mithilfe eines Extraktionsalgorithmus sowie die Berechnung der TF/IDF-Werte für jeden Begriff. Anschließend werden alle TF/IDF-Werte

für jeden Satz summiert. Im nächsten Schritt werden alle Sätze anhand der Summe von TF/IDF eingestuft. Das Kompressionsverhältnis bestimmt die Position des Satzrangs. In dieser Studie wird eine Kompression von 50 % verwendet, was bedeutet, dass die Satzzusammenfassung um 50 % des Originaltextes gekürzt wird. Nach der Auswahl des Satzes wird dessen Berechnung durchgeführt. Die Ähnlichkeit wird mit der Cosinus-Ähnlichkeitsmethode berechnet. Anschließend werden alle Sätze anhand ihrer Cosinus-Ähnlichkeit von der höchsten zur niedrigsten sortiert. Der resultierende Text mit neuer Satzanordnung stellt die finale Zusammenfassung dar.[9]

2.2.5 Pipeline

In dieser Arbeit wird eine Pipeline entwickelt, die die N-Gramm-Analyse verwendet, um Schlagwörter aus einem Text zu extrahieren und mit verschiedenen Ansätzen von Word-Clouds zu visualisieren.[30]

Die vorliegende Arbeit präsentiert eine Anleitung zur Erstellung einer Pipeline mit Python und TF-IDF. Darüber hinaus wird die Relevanz von TF-IDF als Vorverarbeitung beim maschinellen Lernen erörtert. Im Vergleich zur rohen Termhäufigkeit weist TF-IDF in der Regel einen höheren Vorhersagewert auf. Die Gewichtung von Themenwörtern wird erhöht, um die Bedeutung von Wörtern zu erhöhen, während die Gewichtung von hochfrequenten Funktionswörtern verringert wird. Es werden Verfahren zur Vorverarbeitung von Texten vorgestellt, die eine Umformung in die gewünschte Darstellungsform ermöglichen. Zudem werden Methoden zur Interpretation der Ergebnisse des TF-IDF-Verfahrens erörtert.[21]

Die Verarbeitung natürlicher Sprache wirft insbesondere bei der Analyse unüblicher Sprachen wie Griechisch Schwierigkeiten auf. In diesem Beitrag wird ein maschineller Lernansatz für die Bereiche Part-of-Speech-Tagging und Named-Entity-Recognition für die griechische Sprache unter Verwendung von spaCy erarbeitet und evaluiert. [28]

spam-filter (Empfinde das Thema eventuell als zu unpassend)
-Überblick über verfügbare Methoden, Herausforderungen und zukünftige Forschungsrichtungen im Bereich der Spam-Erkennung, Filterung und Eindämmung von SMS-Spam. Dabei werden auch Methodiken der keyword frequency ratio und Herunterbrechung auf keyword components behandelt [34]

—-

2.3 Definitionen und Konzepte: Information Retrieval, Data-Mining

ToDo: Lieber vor dem Kapitel Verwandte Arbeiten packen <–hie

Diese Arbeit beschreibt den Unterschied zwischen Information Filtering und Information Retrieval[3]

g

3 Entwicklung einer klaren Erwartungshaltung

3.1 Beschreibung der Interviews mit Führungskräften zur Identifizierung von Stakeholder-Erwartungen

Zur Beantwortung der Forschungsfragen 2 und 3 dieser Arbeit werden Experteninterviews mit E-Learning Experten durchgeführt. Nachdem für die Beantwortung der Forschungsfrage eins bereits auf die Theorie eingegangen wurde, soll nun diese durch praktische Erfahrungen ergänzt werden. In Forschungsfrage 2 werden die Anforderungen der Praktiker an einen kultursensitiven Leitfaden herausgearbeitet. Um diese Anforderungen angemessen herauszuarbeiten ist eine ausführliche Vorbereitung notwendig. Diese Vorbereitung wird in diesem Kapitel erläutert. Zu Beginn dieses Kapitels wird allgemein auf die qualitative Forschung eingegangen und im Anschluss daran auf die Vorbereitung der Interviews sowie die Vorgehensweise bei der Durchführung der Interviews. Das Kapitel schließt mit einer kurzen Einordnung der Interviewpartner, bezüglich Haupttätigkeitsfeld im E-Learning und der Mitarbeiteranzahl des Unternehmens, ab.
Methodik erklären, siehe Wirtschaftsinformatik Bachelorarbeit (S.34) genau die Schritte der Fragen erklären. Warum diese Reihenfolge
nochmal eine bedarfsmeldung genau erklären und die schritte wie bedarfsmeldung erhalten, gepflegt etc wird erklären

1. Welche Art von Projekten sind typischerweise in Ihrem Unternehmen an der Tagesordnung? Können Sie uns Beispiele für verschiedene Arten von Projekten geben, die *adesso* durchführt?

2. Wie werden Projektbedarfe und -anforderungen innerhalb von *adesso* typischerweise kommuniziert und dokumentiert?

3. Welche Informationen halten Sie in einer Bedarfsmeldung für besonders wichtig oder unverzichtbar?

4. Wie detailliert sollten Projektbeschreibungen Ihrer Meinung nach sein? Sind bestimmte Schlüsselaspekte oder -informationen in jeder Bedarfsmeldung enthalten?

5. Welche Herausforderungen oder Schwierigkeiten sind bei unklaren oder unvollständigen Bedarfsmeldungen aufgetreten?

6. Wer sind die typischen Stakeholder bei der Erstellung von Bedarfsmeldungen und welche Rolle spielen sie?

7. Wie wird die Qualität von Bedarfsmeldungen bei *adesso* bewertet? Gibt es bestimmte Kriterien oder Standards, anhand derer Bedarfsmeldungen beurteilt werden?

8. Wie können Sie die Qualität und Klarheit von Bedarfsmeldungen verbessern?

9. Welche Auswirkungen haben unklare oder fehlende Informationen in Bedarfsmeldungen auf die Effizienz und den Erfolg von Projekten?

10. Wie können Sie sicherstellen, dass die Bedürfnisse und Anforderungen aller relevanten Stakeholder in einer Bedarfsmeldung angemessen berücksichtigt werden?

g

g

g

g

3.2 Analyse der Ergebnisse und Entwicklung einer klaren Erwartungshaltung für die Bedarfsmeldungen

1. Transkription der Interviews:
Falls du die Interviews aufgezeichnet hast, transkribiere sie vollständig und genau. Dadurch hast du eine schriftliche Version der Aussagen der Experten, die du leichter analysieren kannst.

2. Codierung der Daten:
Gehe durch die transkribierten Interviews und markiere oder kodiere relevante Themen, Aussagen oder Muster. Verwende dabei Codes oder Kategorien, die sich auf deine Forschungsfragen beziehen.

3. Thematische Analyse:
Führe eine thematische Analyse durch, indem du die kodierten Daten systematisch durchgehst und nach wiederkehrenden Themen oder Mustern suchst. Identifiziere Gemeinsamkeiten, Unterschiede oder interessante Einsichten, die sich aus den Aussagen der Experten ergeben.

4. Triangulation:
Vergleiche die Ergebnisse der Experteninterviews mit anderen Quellen, wie beispielsweise der Literatur, Fallstudien oder empirischen Daten. Durch die Triangulation kannst du die Glaubwürdigkeit und Validität deiner Ergebnisse erhöhen.

5. Interpretation der Ergebnisse:
Interpretiere die identifizierten Themen oder Muster im Kontext deiner Forschungsfragen und -ziele. Versuche zu verstehen, welche Bedeutung oder Implikationen die Aussagen der Experten für deine Forschung haben könnten.

6. Reflexion und Kritik:
Reflektiere kritisch über die Aussagen der Experten und die gewonnenen Erkenntnisse. Berücksichtige mögliche Einschränkungen oder Bias in den Interviews und betrachte die Ergebnisse aus verschiedenen Perspektiven.

7. Integration in die Gesamtanalyse:
Integriere die Ergebnisse der Experteninterviews in deine Gesamtanalyse deiner Ba-

chelorarbeit. Verknüpfe sie mit anderen Forschungsergebnissen, theoretischen Konzepten oder empirischen Daten, um ein umfassendes Verständnis deines Forschungsthemas zu entwickeln.

8. Darstellung der Ergebnisse:
Präsentiere die wichtigsten Ergebnisse und Erkenntnisse aus den Experteninterviews in deiner Bachelorarbeit. Verwende geeignete Zitate oder Beispiele, um die Aussagen der Experten zu veranschaulichen und deine Argumentation zu unterstützen. [23]
im anhang sind die transskripte wenn man nicht ne größere anzahl an infos hat gucken ob man das halb automatisch evaluieren. Vielleicht kategorisieren. Infos die wichtig sind gucken ob die dann auch nach dem preprocessing drin sind. Regressive tests schreiben.
transformation von bedarfsmeldung zu guter bedarfsmeldung, was ist der fokus von der bedarfsmeldung, wie gut machen die ansätze das, und muss man das dann noch weiter verarbeiten, haben wir alles was wir brauchen mit nur einem algorithmus, inferenz falls parameter fehlt, gibt es einen der alles löst
Fragen in das proposal aufnehmen, führungskraft vorher fragen ob die fragen nice sind.

g

1. Wer sind die typischen Stakeholder bei der Erstellung von Bedarfsmeldungen und welche Rolle spielen sie?
2. Welche Art von Projekten sind typischerweise in Ihrem Unternehmen an der Tagesordnung? Können Sie uns Beispiele für verschiedene Arten von Projekten geben, die adesso durchführt?
3. Wie werden Projektbedarfe und -anforderungen innerhalb von adesso typischerweise kommuniziert und dokumentiert?
4. Welche Informationen halten Sie in einer Bedarfsmeldung für besonders wichtig oder unverzichtbar?
5. Wie detailliert sollten Projektbeschreibungen Ihrer Meinung nach sein? Sind bestimmte Schlüsselaspekte oder -informationen in jeder Bedarfsmeldung enthalten?
6. Wie wird die Qualität von Bedarfsmeldungen bei adesso bewertet? Gibt es bestimmte Kriterien oder Standards, anhand derer Bedarfsmeldungen beurteilt werden?
7. Wie können Sie die Qualität und Klarheit von Bedarfsmeldungen verbessern?
8. Welche Herausforderungen oder Schwierigkeiten sind bei unklaren oder unvollständigen Bedarfsmeldungen aufgetreten?
9. Welche Auswirkungen haben unklare oder fehlende Informationen in Bedarfsmeldungen auf die Effizienz und den Erfolg von Projekten?
10. Wie können Sie sicherstellen, dass die Bedürfnisse und Anforderungen aller relevanten Stakeholder in einer Bedarfsmeldung angemessen berücksichtigt werden?

g

4 Analyse der Techniken des Information Retrieval und Data-Mining

4.1 Beschreibung der untersuchten Techniken und Ansätze

4.1.1 TF-IDF

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 gelten als potenzielle Schlüsselwörter.
[2][31]

4.1.2 Text-Ranking-Algorithmen

Text-Ranking-Algorithmen: Text-Ranking-Algorithmen wie TextRank oder YAKE (Yet Another Keyword Extractor) verwenden graphenbasierte Methoden, um Schlüsselwörter in einem Text zu identifizieren. Die Algorithmen bewerten die Wichtigkeit von Wörtern basierend auf ihrer Verbindung zu anderen Wörtern im Text und extrahieren Schlüsselwörter entsprechend ihrer Rangfolge.
[25][36][29]

4.1.3 N-Gramm

N-Gramm-Analyse: N-Gramme sind Sequenzen von N aufeinanderfolgenden Wörtern in einem Text. Durch die Analyse von N-Grammen können häufig vorkommende Phrasen oder Begriffe identifiziert werden, die potenzielle Schlüsselwörter darstellen.
[30]

4.1.4 POS-Tagging

Part-of-Speech (POS) Tagging: POS-Tagging wird genutzt, 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.
[19][27]

4.1.5 Named Entity Recognition

Named Entity Recognition (NER) [24] [26][28]

4.1.6 Regelbasierte Ansätze

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.

4.1.7 Hybride Ansätze

Hybride Ansätze: Hybride Ansätze kombinieren verschiedene Methoden und Techniken, um eine genauere Extraktion von Schlüsselwörtern zu ermöglichen. (Z.B. Kombination aus TF-IDF-Gewichtung und Text-Ranking-Algorithmen verwendet).[9]

4.1.8 Data-Mining

Data-Mining: [16][15]

4.1.9 preprocessing

preprocessing: [13]

4.1.10 data-fusion

data-fusion: [11] [12] [4]

g

g

g

4.2 Bewertung und Auswahl der besten Ansätze für die Extraktion relevanter Inhalte aus Bedarfsmeldungen

g

5 Konzeptionierung und Implementierung der Vorverarbeitung

5.1 Beschreibung des entwickelten Vorverarbeitungsmodells

basierend auf den ausgewählten Techniken
-auch translate erwähnen. Wichtig damit die meisten ansätze gut funktionieren

g

g

g

g

5.2 Details zur Implementierung der Pipeline in Python

für die effiziente Verarbeitung von Bedarfsmeldungen

g

6 Evaluierung des entwickelten Systems

6.1 Beschreibung des verwendeten Datensatzes und der Evaluierungsmethodik

überlegung ob tfidf unterschied macht alle bedarfsmeldungen mit einer zu vergleichen und daraus wichtige wörter identifizieren oder eine für sich alleine reicht.
gucken was tokenisierung wirklich macht

6.2 Präsentation und Diskussion der Ergebnisse

g

g

g

Zeit und Leistung Übersicht

g

6.3 Vergleich des Systems mit einem Large Language Model-Ansatz

g

g

g

6.4 Analyse von Abweichungen, Ähnlichkeiten und Verbesserungspotenzialen des Systems

g

7 Zusammenfassung und Ausblick

ergebnis der arbeit: diese modelle in der reihenfolge kommen am nähesten an die bedarfsmeldung

Ausblick

die keyword extraction auch für die profile nutzen

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig angefertigt und mich keiner fremden Hilfe bedient sowie keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen, die wörtlich oder sinngemäß veröffentlichten oder nicht veröffentlichten Schriften und anderen Quellen entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

7.1 Erklärung zu eingesetzten Hilfsmitteln

1. Korrekturservice der Fachhochschule bzw. des Fachbereichs genutzt:
   - ☐ Ja
   - ☒ Nein
2. Einsatz eines externen (kommerziellen) Korrekturservice:
   - ☐ Ja
   - ☒ Nein
3. Folgende Personen haben die Arbeit zusätzlich Korrektur gelesen:
   - ...
4. Nutzung von Sprachmodellen für die Texterstellung (z.B. ChatGPT), wenn ja, welche und in welchen Abschnitten:
   - ☐ Ja
   - ☒ Nein
5. Sprachübersetzungstools (z.B. Google Übersetzer, DeepL), wenn ja, welche und in welchen Abschnitten:
   - ☒ Ja
   - ☐ Nein

- DeepL, Im Kapitel Literaturüberblick für das bessere Verständnis der Literatur

6. Einsatz von Software zur Sprachkorrektur (z.B. Grammarly), wenn ja, welche und in welchen Abschnitten:

   ☐ Ja

   ☐ Nein

   - ...

7. Einsatz anderer Hilfsmittel:

   -

8. Ich stimme dem möglichen Einsatz von Software zur Plagiatserkennung zu:

   ☒ Ja

   ☐ Nein

Ich bestätige, dass obige Aussagen vollständig und nach bestem Wissen ausgefüllt wurden.

Dortmund, den 24. April 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] T. Bohne und U. M. Borghoff, "Data fusion: Boosting performance in keyword extraction," in *2013 20th IEEE International Conference and Workshops on Engineering of Computer Based Systems (ECBS)*, IEEE, 2013, S. 166–173.

[5] R. Burke, A. Felfernig und M. H. Göker, "Recommender systems: An overview," *Ai Magazine*, Jg. 32, Nr. 3, S. 13–18, 2011.

[6] G. Chartron und G. Kembellec, "General introduction to recommender systems," *Recommender Systems*, S. 1–23, 2014.

[7] M. Chiny, M. Chihab, O. Bencharef und Y. Chihab, "LSTM, VADER and TF-IDF based hybrid sentiment analysis model," *International Journal of Advanced Computer Science and Applications*, Jg. 12, Nr. 7, 2021.

[8] 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.

[9] R. Darmawan und R. S. Wahono, "Hybrid Keyword Extraction Algorithm and Cosine Similarity for Improving Sentences Cohesion in Text Summarization," *Journal of Intelligent Systems*, Jg. 1, Nr. 2, S. 109–114, 2015.

[10] Z. Dong, Z. Wang, J. Xu, R. Tang und J. Wen, "A brief history of recommender systems," *arXiv preprint arXiv:2209.01860*, 2022.

[11] A Famili, W.-M. Shen, R. Weber und E. Simoudis, "Data preprocessing and intelligent data analysis," *Intelligent data analysis*, Jg. 1, Nr. 1, S. 3–23, 1997.

[12] D Frank Hsu und I. Taksa, "Comparing rank and score combination methods for data fusion in information retrieval," *Information retrieval*, Jg. 8, Nr. 3, S. 449–480, 2005.

[13] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez und F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, Jg. 1, Nr. 1, S. 1–22, 2016.

[14] 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.

[15] N. Jain und V. Srivastava, "Data mining techniques: a survey paper," *IJRET: International Journal of Research in Engineering and Technology*, Jg. 2, Nr. 11, S. 2319–1163, 2013.

[16] S. Jun Lee und K. Siau, "A review of data mining techniques," *Industrial Management & Data Systems*, Jg. 101, Nr. 1, S. 41–46, 2001.

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

[18] P. Kroha, "Preprocessing of requirements specification," in *Database and Expert Systems Applications: 11th International Conference, DEXA 2000 London, UK, September 4–8, 2000 Proceedings 11*, Springer, 2000, S. 675–684.

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

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

[21] M. Lavin, "Analyzing documents with TF-IDF," 2019.

[22] J. Lu, Q. Zhang und G. Zhang, *Recommender systems: advanced developments*. World Scientific, 2020.

[23] M. Maguire und N. Bevan, "User requirements analysis: a review of supporting methods," in *IFIP World Computer Congress, TC 13*, Springer, 2002, S. 133–148.

[24] A. Mansouri, L. S. Affendey und A. Mamat, "Named entity recognition approaches," *International Journal of Computer Science and Network Security*, Jg. 8, Nr. 2, S. 339–344, 2008.

[25] 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.

[26] D. Nadeau und S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, Jg. 30, Nr. 1, S. 3–26, 2007.

[27] 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.

[28] E. Partalidou, E. Spyromitros-Xioufis, S. Doropoulos, S. Vologiannidis und K. Diamantaras, "Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy," in *IEEE/WIC/ACM International Conference on Web Intelligence*, 2019, S. 337–341.

[29] 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.

[30] S. Pirk, "Implementierung und Visualisierung N-Gramm-basierter Word-Clouds," B.S. thesis, 2019.

[31] 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.

[32] S. Reddy, S. Nalluri, S. Kunisetti, S Ashok und B Venkatesh, "Content-based movie recommendation system using genre correlation," in *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018, Volume 2*, Springer, 2019, S. 391–397.

[33] F. Ricci, *Recommender Systems: Models and Techniques.* 2014.

[34] 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.

[35] V Suhasini und N Vimala, "A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data," *Turkish Journal of Computer and Mathematics Education*, Jg. 12, Nr. 6, S. 5710–5723, 2021.

[36] 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.