Proposal zur 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 21. Februar 2024

Motivation

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 Fachkräften 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.

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.

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.

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 *TF-IDF*, *Text-Ranking-Algorithmen*, *N-Gramm-Analyse*, *POS-Tagging*, *Named Entity Recognition*, Regelbasierte Ansätze und Hybride Ansätze 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 dieser Modelle 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 in die Pipeline eingefügt. Dabei wird überprüft, inwiefern das Ergebnis der definierten Erwartungshaltung entspricht. Mit Hilfe einer manuellen Überprüfung werden Abweichungen, Ähnlichkeiten und Anpassungen analysiert, um Erkenntnisse darüber zu gewinnen, wie das System inhaltlich abschneidet und welche Techniken allein oder in Kombination mit mehreren Ansätzen die wichtigsten Informationen filtert. Da die Dauer eine entscheidende Rolle spielt, werden auch Zeit und Leistung gemessen.

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 Data-Mining-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

Aufbau der Arbeit

1 Einleitung **8**
1.1 Problemstellung 9
1.2 Ziele und Ergebnisse der Arbeit 10
1.3 Aufbau der Arbeit 10

2 Literaturüberblick **11**
2.1 Definitionen und Konzepte: Information Retrieval, Data-Mining, Bedarfsmeldungen 13

3 Entwicklung einer klaren Erwartungshaltung **14**
3.1 Beschreibung der Interviews mit Führungskräften zur Identifizierung von Stakeholder-Erwartungen 14
3.2 Analyse der Ergebnisse und Entwicklung einer klaren Erwartungshaltung für die Bedarfsmeldungen 15

4 Analyse der Techniken des Information Retrieval und Data-Mining **17**
4.1 Beschreibung der untersuchten Techniken und Ansätze 17
4.2 Bewertung und Auswahl der besten Ansätze für die Extraktion relevanter Inhalte aus Bedarfsmeldungen 18

5 Konzeptionierung und Implementierung der Vorverarbeitung **19**
5.1 Beschreibung des entwickelten Vorverarbeitungsmodells basierend auf den ausgewählten Techniken 19
5.2 Details zur Implementierung der Pipeline in Python für die effiziente Verarbeitung von Bedarfsmeldungen 19

6 Evaluierung des entwickelten Systems **20**
6.1 Beschreibung des verwendeten Datensatzes und der Evaluierungsmethodik 20
6.2 Präsentation und Diskussion der Ergebnisse basierend auf den Metriken Precision, Recall und F1-Score 20
6.3 Analyse von Abweichungen, Ähnlichkeiten und Verbesserungspotenzialen des Systems 20

7 Zusammenfassung und Ausblick **21**
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,
was muss ich jetzt machen: gucken wie ich das inhaltlich genau machen will, also pipeline genauch checken, quellen von der bachelorarbeit checken

1 Einleitung

1.1 Problemstellung

1.2 Ziele und Ergebnisse der Arbeit

1.3 Aufbau der Arbeit

2 Literaturüberblick

-Einschätzen der Fähigkeiten, Talente und des Fachwissens der Mitarbeiter
-In diesem Papier wird ein Ansatz beschrieben, um aus Unternehmensdaten und den digitalen Fußabdrücken der Mitarbeiter Informationen zu gewinnen.
-Beurteilung des Fachwissens eines Mitarbeiters in einem breiten Bereich wie cloud computing oder cybersecurity
-Auf einer hohen Ebene lässt sich der Ansatz der Informationsbeschaffung und -fusion wie folgt beschreiben: Es wird eine Liste von Suchbegriffen erstellt, die sich auf das breite Fachgebiet beziehen.
-Die Suche wird nach jedem dieser Abfragebegriffe durchgeführt, um Beweise für Mitarbeiter und Datenquellen zu finden. Die verschiedenen Beweisstücke werden miteinander verschmolzen, gewichtet und nach der Abfrage sortiert. Die Mitarbeiter werden nach Datenquelle gewichtet und möglicherweise auf andere Weise bewertet, um einen einzigen Ordinalwert (sehr niedrig, niedrig, moderat, etwas, begrenzt) für ihr Fachwissen in diesem breiten Bereich zu erhalten.[11]

information filtering
-Informationen für seine Benutzer betreffend der Anwender in Bezug auf ihre Interessengebiete zu reduzieren -Dazu werden nicht relevante Dokumente aus einem Strom von Informationen entfernt, sodass den Anwendern nur relevante Dokumente präsentiert werden. -Ein Teil der Arbeit beschäftigt sich mit der der Informationsfilterung und mögliche Filterungsvarianten werden vorgestellt. Die Arbeit konzentriert sich auf die inhaltsbasierte Filterung von Textdokumenten und identifizieren Informationsfilterung als einen Spezialfall der Textklassifikation. -Überblick über gängige Methoden. Anschließend werden bekannte Filterungsprojekte kurz vorgestellt, bevor verwandte Aufgaben verglichen werden. [17]
preprocessing [1] -Wege und Schritte zur Aufbereitung von Datensätzen -Arbeit umfasst Data-Mining Vorverarbeitung um Qualität der Daten zu verbessern -Wichtiger Schritt um Effizienz zu verbessern
——- spam-filter
-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 [28]

—- In diesem Beitrag werden Studien zu Technologien vorgestellt, die für die Suche und das Abrufen von Informationen im Web nützlich sind. Es wird aufgezeigt, dass Information Retrieval und Ranking im Web-Kontext anders Funktioniert als in einer statischen Datenbank. [14]

Die Kombination von verschiedenen Textdarstellungen und Suchstrategien ist zu einer Standardtechnik geworden, um die Effektivität der Informationsbeschaffung zu verbessern.[6]

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.[26]

python pipeline mit python und tf-idf. Beschreibt warum TF-IDF häufig in als Vorverarbeitung beim maschinellen Lernen eingesetzt wird. Hat in der Regel einen höheren Vorhersagewert als rohe Termhäufigkeit. Die Gewichtung von Themenwörtern wird erhöt„ um die Bedeutung von Wörtern zu erhöhen, während die Gewichtung von hochfrequenten Funktionswörtern verringert wird.[18]

Kombination drei Ansätze Ansätze. Unter anderem auch TF-IDF. Kombinieren mit einem sogenannten CLASSIFIER Model. Das Klassifikationsmodell bezieht sich direkt auf die Ergebnisse der Modelle LSTM, VADER und TFIDF, die jeweils drei Eingaben liefern. Die Werte dieser Eingaben liegen im Bereich von [0,1]. Die Ausgabe des Klassifikationsmodells ist binär und liefert eine Vorhersage der Stimmung des vollständigen Textes der Modelleingabe (positiv oder negativ).[5]

Das erste vorverarbeitete Dokument wird mithilfe eines Extraktionsalgorithmus analysiert und anschließend wird für jeden Begriff TF/IDF berechnet. Danach werden alle TF/IDF-Begriffe 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 seine Berechnung durchgeführt. 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 ist die endgültige Zusammenfassung.[7]

Kombination aus TD-IDF und N-Gram. Um Fake news heraus zu filtern[29]

Named ENtity Recognition mit POS-Tagger Implementierung mit Spacy für die griechische Sprache[24]

Preprocessing von Softwareanforderungen. Generierung aus Text Anforderungen zu diversen Diagrammen etc [15]

2.1 Definitionen und Konzepte: Information Retrieval, Data-Mining, Bedarfsmeldungen

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

3 Entwicklung einer klaren Erwartungshaltung

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

- 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?
- Wie werden Projektbedarfe und -anforderungen innerhalb von *adesso* typischerweise kommuniziert und dokumentiert?
- Welche Informationen halten Sie in einer Bedarfsmeldung für besonders wichtig oder unverzichtbar?
- Wie detailliert sollten Projektbeschreibungen Ihrer Meinung nach sein? Sind bestimmte Schlüsselaspekte oder -informationen in jeder Bedarfsmeldung enthalten?
- Welche Herausforderungen oder Schwierigkeiten sind bei unklaren oder unvollständigen Bedarfsmeldungen aufgetreten?
- Wer sind die typischen Stakeholder bei der Erstellung von Bedarfsmeldungen und welche Rolle spielen sie?
- Wie wird die Qualität von Bedarfsmeldungen bei *adesso* bewertet? Gibt es bestimmte Kriterien oder Standards, anhand derer Bedarfsmeldungen beurteilt werden?
- Wie können Sie die Qualität und Klarheit von Bedarfsmeldungen verbessern?
- Wie können Sie die Qualität und Klarheit von Bedarfsmeldungen verbessern?
- Welche Auswirkungen haben unklare oder fehlende Informationen in Projektbeschreibungen auf die Effizienz und den Erfolg von Projekten?
- Wie können Sie sicherstellen, dass die Bedürfnisse und Anforderungen aller relevanten Stakeholder in einer Bedarfsmeldung angemessen berücksichtigt werden?

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 Bachelorarbeit. 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. [19]

4 Analyse der Techniken des Information Retrieval und Data-Mining

4.1 Beschreibung der untersuchten Techniken und Ansätze

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][27]

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.
[21][30][25]

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.
[26]

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.
[16][23]

Named Entity Recognition (NER) [20] [22][24]

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.

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).[7]

Data-Mining: [13][12]
preprocessing: [10]
data-fusion: [8] [9] [4]

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

5 Konzeptionierung und Implementierung der Vorverarbeitung

5.1 Beschreibung des entwickelten Vorverarbeitungsmodells basierend auf den ausgewählten Techniken

5.2 Details zur Implementierung der Pipeline in Python für die effiziente Verarbeitung von Bedarfsmeldungen

6 Evaluierung des entwickelten Systems

6.1 Beschreibung des verwendeten Datensatzes und der Evaluierungsmethodik

6.2 Präsentation und Diskussion der Ergebnisse basierend auf den Metriken Precision, Recall und F1-Score

6.3 Analyse von Abweichungen, Ähnlichkeiten und Verbesserungspotenzialen des Systems

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

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 21. 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] 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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

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

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

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.