Titel TBD.

Optionele ondertitel.

Thomas Vanderveken.

Scriptie voorgedragen tot het bekomen van de graad van
Professionele bachelor in de toegepaste informatica

Promotor: TBD
Co-promotor: TBD
Academiejaar: 2023–2024
Derde examenperiode

Departement IT en Digitale Innovatie .

HO
GENT

Woord vooraf

Samenvatting

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

Lijst van codefragmenten

1

Inleiding

Regelgevende filings door institutionele beleggers, zoals pensioenfondsen en vermogensbeheerders, bieden belangrijke inzichten in marktpatronen en beleggingsstrategie. De 13F-bestanden die worden ingediend bij de Amerikaanse Securities and Exchange Commission (SEC) zijn een van de belangrijkste bronnen van deze informatie. Deze registraties geven informatie over de bezittingen van institutionele beleggers, waardoor ze cruciaal zijn voor het doen van financieel onderzoek en het analyseren van beleggingen. 13F-bestanden van vóór 2013 leveren echter aanzienlijke problemen op vanwege hun variabele vormen en structuren, die de menselijke verwerking en analyse complexer maken.

De opkomst van geavanceerde AI-technologie biedt een potentiële kans om deze problemen aan te pakken. Natural Language Processing (NLP) en Machine Learning (ML) bieden geavanceerde technieken voor het extraheren en organiseren van gegevens uit tekst zonder vooraf gedefinieerde structuur. Door gebruik te maken van deze technologieën is het mogelijk om het proces van het standaardiseren en combineren van eerdere 13F aanvragen in een goed georganiseerde relationele database te automatiseren, waardoor de toegang en het gebruik wordt verbeterd.

Het doel van dit proefschrift is het creëren van een proof-of-concept toepassing die Natural Language Processing (NLP) en Machine Learning (ML) technieken gebruikt om 13F aanvragen van vóór 2013 te uniformeren en te integreren in een relationele database. Het voorgestelde systeem is gericht op het stroomlijnen van het gegevensextractieproces door de deponeringen automatisch om te zetten in een gestandaardiseerd formaat met een hoge efficiëntie en nauwkeurigheid. Dit zou niet alleen de analyse van financiële gegevens uit het verleden optimaliseren, maar ook het werk en de kosten verminderen die gepaard gaan met handmatige gegevensverwerking.

Bovendien zouden de gestandaardiseerde gegevens het begrip van investeringspatronen uit het verleden verbeteren en het creëren van voorspellingsmodellen

ondersteunen. Het onderzoek zal beginnen met een uitgebreide literatuurstudie om de meest efficiënte Natural Language Processing (NLP) en Machine Learning (ML) strategieën voor deze specifieke onderneming te bepalen. Daarna zal een proof-of-concept toepassing worden gecreëerd en beoordeeld op nauwkeurigheid, efficiëntie en bruikbaarheid.
Deze inleiding geeft een beknopt overzicht van de redenen, doelen en het belang van het onderzoek. Dit werk wil een nuttige bijdrage leveren aan de analyse van financiële gegevens en onderzoekers en analisten een nuttig hulpmiddel bieden door de moeilijkheden aan te pakken die gepaard gaan met het verwerken van oudere 13F-bestanden.

1.1. Probleemstelling

13F meldingen van de SEC voor 2013, zijn belangrijke bestanden voor financieel onderzoek, ze bevatten namelijk data over de stocks dat investment managers beheren. Maar deze zijn vaak inconsistent in opmaak en moeilijker toegankelijk, wat manuele analyse bemoeilijkt. Er ontbreekt namelijk een geautomatiseerd systeem om deze gegevens te standaardiseren en in een databank te integreren. Dit bemoeilijkt de opportuniteiten voor diepgaande analyses en het verkrijgen van inzichten in beleggingstrends. Dit onderzoek gaat opzoek naar hoe AI-technologieën zoals NLP en ML, ingezet kunnen worden om deze meldingen te extraheren, te structuren en te integreren in een databank, wat als gevolg het gebruik en de toegankelijkheid van historische financiële gegevens te verbeteren.

1.2. Onderzoeksvraag

Hoe kunnen AI-technologieën zoals Natural Language Processing (NLP) en Machine Learning (ML) effectief worden toegepast om 13F-meldingen van de SEC van vóór 2013 te standaardiseren en te integreren in een gestructureerde databank, zodat de historische gegevens efficiënter kunnen worden geanalyseerd en vergeleken?

1.3. Onderzoeksdoelstelling

Het hoofddoel van dit onderzoek is het ontwikkelen van een geautomatiseerde methode die gebruikmaakt van AI-technologieën, zoals NLP en ML, om de data uit de 13F meldingen van voor 2013 te extraheren, standaardiseren en te integreren in een relationele databank. Dit moet leiden tot een efficiëntere en meer accurate extractie van gegevens uit deze documenten, waardoor de toegankelijkheid en bruikbaarheid van de data voor financieel onderzoek en investeringsanalyse aanzienlijk worden verbeterd.

1.4. Opzet van deze bachelorproef

Het verdere verloop van deze bachelorproef is opgebouwd als volgt:
In Hoofdstuk 2 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomein, op basis van een literatuurstudie.
In Hoofdstuk 3 wordt de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.
In Hoofdstuk 3 wordt de proof-of-concept besproken. De inhoud omvat de ingewikkelde technische specificaties, structuur en tools, samen met de functionele elementen zoals de modellen en de databank.
In Hoofdstuk 4, tenslotte, wordt de conclusie gegeven en een antwoord geformuleerd op de onderzoeksvragen. Daarbij wordt ook een aanzet gegeven voor toekomstig onderzoek binnen dit domein.

2

Stand van zaken

De Securities and Exchange Commission (SEC) vereist dat institutionele vermogensbeheerders een kwartaalrapport indienen dat bekend staat als Form 13F als ze zeggenschap hebben over $100 miljoen of meer in sectie 13(f) effecten. Sectie 13(f) van de Securities Exchange Act van 1934 verplicht de openbaarmaking van effectenbezit door grote institutionele beleggers om de transparantie te vergroten. In 1975 implementeerde het Congres deze bepaling om de toegankelijkheid van informatie over de investeringsactiviteiten van deze bedrijven te verbeteren. De bedoeling was om het vertrouwen van beleggers in de integriteit van de effectenmarkten in de Verenigde Staten te vergroten door middel van een openbaarmakingsprogramma(**SECform13** Formulier 13F biedt een uitgebreid overzicht van de aandelenbeleggingen van prominente beleggingsmaatschappijen wereldwijd en is een zeer belangrijk hulpmiddel voor analisten, onderzoekers en beleggers die inzicht willen krijgen in markttrends en de beleggingsbenaderingen van belangrijke marktspelers. Het onverwerkte tekstformaat waarin deze inzendingen worden aangeleverd, vormt echter een aanzienlijke belemmering voor effectieve gegevensextractie en -analyse, vooral voor inzendingen van vóór 2013. Vóór 2013 ontbrak het bij 13F-meldingen vaak aan standaardisatie en systematische opmaak, wat nu wel gebruikelijk is bij recentere aanmeldingen.

Kunstmatige intelligentie (AI) en machine learning (ML) technologieën hebben de extractie en organisatie van gegevens uit ongestructureerde tekst de afgelopen jaren aanzienlijk veranderd. Geavanceerde methodologieën zoals Natural Language Processing (NLP) en deep learning modellen vergemakkelijken de omzetting van tekstuele 13F aanvragen in gestructureerde datasets die geschikt zijn voor grondige analyse en studie. Standaardisatie is cruciaal voor historische gegevens, omdat het ontbreken van uniformiteit geautomatiseerde verwerking kan bemoeilijken. Door gebruik te maken van deze technologieën kunnen we zowel huidige als vroegere 13F aanvragen omzetten in georganiseerde gegevens, die vervolgens

kunnen worden opgeslagen in databases, waardoor patronen eenvoudiger kunnen worden opgehaald, gevisualiseerd en geanalyseerd.

Het doel van deze literatuurstudie is het onderzoeken en beoordelen van de verschillende Artificial Intelligence (AI) en Machine Learning (ML) technieken die kunnen worden gebruikt om gegevens uit 13F-meldingen van voor 2013 te extraheren, te organiseren en op te slaan. Het doel van het onderzoek is het bepalen van de meest efficiënte methoden om de ongeorganiseerde inhoud van deze documenten om te zetten in een gestructureerd formaat dat geschikt is voor analyse en opslag in een database. Dit houdt in dat er een vergelijkend onderzoek wordt gedaan naar verschillende kunstmatige intelligentie methodologieën, zoals Natural Language Processing (NLP) en Deep Text mining, en dat bepaalde tools zoals NLTK en SpaCy worden geëvalueerd. De evaluatie zal ook de integratie van gestructureerde gegevens in databasemanagementsystemen (DBMS) onderzoeken, om te garanderen dat de geëxtraheerde gegevens gemakkelijk beschikbaar zijn voor later onderzoek en analyse. Het doel van deze evaluatie is om een uitgebreide kennis te krijgen van de meest effectieve procedures en technologie voor het verwerken van 13F-meldingen.

2.1. Wat zijn 13F meldingen

Het doel van dit gedeelte is om een beknopte introductie te geven tot 13F deponeringen, met de nadruk op de structuur, de informatie die ze bevatten en hun doel.

2.1.1. Definitie en doel

Volgens (**SECform13F2024**) zijn 13F-meldingen verplichte wettelijke documenten die de Amerikaanse Securities and Exchange Commission (SEC) vereist onder Sectie 13(f) van de Securities Exchange Act van 1934. Deze deponeringen worden gebruikt om de portefeuilles van institutionele beleggingsbeheerders te rapporteren. Het belangrijkste doel van 13F meldingen is om duidelijkheid en openheid te bieden over de beleggingsactiviteiten van belangrijke institutionele beleggers. Deze vereiste vergemakkelijkt het toezicht op beleggingsposities van verschillende instellingen, zoals beleggingsfondsen, pensioenfondsen en andere belangrijke beleggingsbeheerders, door het publiek en regelgevende instanties.

2.1.2. Belangrijke kenmerken

Het doel van dit gedeelte is om een beknopte introductie te geven tot 13F deponeringen, met de nadruk op de structuur, de informatie die ze bevatten en hun doel.

Vereisten voor rapportage:
Frequentie: Institutionele beleggingsbeheerders met minimaal $100 miljoen aan beheerd vermogen moeten elk kwartaal een 13F-meldingen indienen. Inhoud: De rapporten bevatten uitgebreide informatie over het aandelenbezit van de instelling, waaronder de naam, CUSIP-nummer, het aantal aandelen en de marktwaarde.

Omvang van de informatie:
Aandelenbezit: De openbaarmakingen concentreren zich voornamelijk op aandelen, terwijl andere soorten activa zoals obligaties, derivaten en private equity worden uitgesloten. Disclaimer: Elk bestand biedt een kort overzicht van de aandelenportefeuille van de instelling aan het einde van de rapportageperiode, wat een waardevol inzicht geeft in hun investeringsmethoden.

Opmaak en toegankelijkheid:
Vanaf 2013 is het verplicht om alle 13F-meldingen elektronisch in te dienen via het EDGAR-systeem van de SEC. De elektronische deponeringen zijn gemakkelijk toegankelijk voor het publiek, wat transparantie garandeert en het bestuderen van het materiaal vergemakkelijkt. Papieren meldingen (vóór 2013): Vóór 2013 werd een aanzienlijk aantal 13F-meldingen op papier ingediend, wat resulteerde in een moeilijker en tijdrovender proces om toegang te krijgen tot de informatie en deze te analyseren. Deze dossiers moesten vaak door mensen in systemen worden ingevoerd voor verdere verwerking.
Illustraties van 13F indieningsformaten:

- Voorbeeld van papieren indiening vóór 2013:
  - Dit is een gedigitaliseerde afbeelding van een standaard 13F-meldingen dat traditioneel op papier werd ingediend voordat elektronische indiening verplicht werd. Dit formaat omvatte handmatig geschreven of getypte informatie over de activa van de instelling.
  - Het analyseren van papieren bestanden leverde aanzienlijke problemen op bij het extraheren en analyseren van gegevens, waardoor het gebruik van handmatige gegevensinvoer- en validatieprocedures noodzakelijk werd.
- Na het jaar 2013, als voorbeeld van elektronische archivering:
  - Vanaf 2013 werd het elektronische formaat gestandaardiseerd, waardoor onmiddellijke toegang tot en analyse van deponeringen vanuit het EDGAR-systeem mogelijk werd. Hieronder ziet u een momentopname van een elektronische 13F indiening.
  - Het gebruik van elektronische deponeringen heeft de efficiëntie van gegevensanalyse aanzienlijk verbeterd door geautomatiseerde extractie en vereenvoudigde aggregatie van financiële gegevens mogelijk te maken.

Historisch onderzoek van deponeringen van vóór 2013:

- Hoewel deponeringen van na 2013 gemakkelijk beschikbaar zijn en door machines verwerkt kunnen worden, bieden deponeringen van voor 2013 rijke historische informatie die essentieel is voor het doen van diepgaande marktanalyses en onderzoek over een langere periode. Niettemin moeten de eerdere dossiers meer bewerkingen ondergaan om de gegevens te converteren en te organiseren voor analyse.
- Er ontstaan uitdagingen bij het extraheren van gegevens wanneer geprobeerd wordt om informatie uit papieren bestanden van voor 2013 te halen. Dit proces vereist het gebruik van OCR-technologie (Optical Character Recognition), handmatige gegevensinvoer of methoden voor gegevensinvoer via crowdsourcing. Deze gebreken en inconsistenties moeten in elke analyse worden meegenomen.
- De periode voorafgaand aan het elektronische indieningsmandaat in 2013 markeerde een belangrijke verschuiving in de rapportage en toegankelijkheid van financiële gegevens. Deze periode toonde de vooruitgang in financiële rapportagepraktijken en het toenemende belang van digitale gegevensverwerking in financiële analyses.

Samenvatting van de structuur en het gebruiksgemak De overgang van papieren naar elektronische 13F-meldingen betekende een aanzienlijke verbetering in de beschikbaarheid en gebruiksvriendelijkheid van financiële gegevens voor zowel beleggers als wetenschappers. Desondanks heeft het bestaan van bestanden die vóór 2013 zijn gemaakt specifieke moeilijkheden en voordelen voor het uitvoeren van historische analyses. Het is van cruciaal belang om de verschillende formaten en hun gevolgen voor de extractie en analyse van gegevens te begrijpen.

CHAPTER UNDER REVIEW TODO

2.2. AI en Machine learning in financiële data extractie

Dit hoofdstuk onderzoekt de significante invloed van kunstmatige intelligentie (AI) en machinaal leren (ML) op het ophalen van financiële gegevens, met name op 13F-meldingen van vóór 2013. Deze tekst presenteert de basisprincipes van kunstmatige intelligentie (AI) en machinaal leren (ML), onderzoekt hoe ze worden gebruikt bij het extraheren van financiële gegevens en analyseert praktijkvoorbeelden die de doeltreffendheid ervan aantonen. Het hoofdstuk bespreekt ook specifieke moeilijkheden in verband met gegevenskwaliteit, de veranderende financiële terminologie en de noodzaak om AI-modellen aan te passen en aan te passen om de complexiteit van historische financiële documentatie effectief te beheren. 2.1 Introduction to AI and Machine Learning 2.1.1 Overview of AI and Machine Learning Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized the way

data is processed and analyzed, particularly in fields that deal with large volumes of unstructured or semi-structured data, such as finance. AI involves the development of systems capable of performing tasks that typically require human intelligence, while ML, a subset of AI, focuses on the creation of algorithms that can learn from data and make predictions or decisions.
2.1.2 Applications of AI in Financial Data Extraction In the financial sector, AI is used extensively for tasks like automated trading, risk assessment, and regulatory compliance. Specifically, AI plays a crucial role in the extraction of data from financial documents. Techniques such as Natural Language Processing (NLP) and Optical Character Recognition (OCR) are essential for interpreting and converting textual information from forms like 13F meldingen into structured data that can be analyzed.
2.1.3 Case Studies in AI-Driven Data Extraction There are numerous examples where AI has been applied successfully to extract data from financial documents. For instance, AI-driven systems have been developed to analyze 10-K and 10-Q meldingen, which share similarities with 13F forms in terms of structure and complexity. These case studies highlight the potential for AI to automate and improve the accuracy of data extraction processes, particularly for regulatory meldingen.
2.2 Challenges in Applying AI to Pre-2013 13F Forms 2.2.1 Data Quality and Format Issues Pre-2013 13F forms present unique challenges due to their non-standardized formats and the potential for poor data quality, such as low-resolution scans or handwritten annotations. These factors can significantly hinder the effectiveness of AI tools, particularly OCR systems, which rely on clear and consistent text patterns.
2.2.2 Evolution of Financial Language The financial terminology used in older 13F forms may differ from modern usage, posing additional challenges for NLP models that are trained on contemporary datasets. This evolution of language requires the customization of AI tools to accurately interpret and extract relevant information from these historical documents.
2.2.3 Customization and Adaptation of AI Models Addressing these challenges requires adapting existing AI models. For example, OCR models might need retraining to recognize outdated fonts or handwritten text, while NLP models may require fine-tuning to understand the specific financial terminology of the time. Additionally, hybrid approaches that combine rule-based systems with machine learning may be necessary to handle the inconsistencies and complexities of these older documents.

2.3. Text mining en gerelateerde technieken

Text mining, of ook bekend tekstdatamining, is de procedure om ongestructureerde tekst om te zetten in een gestructureerd formaat om significante patronen te ont-

dekken en nieuwe inzichten te verwerven. Text mining maakt de analyse van uitgebreide tekstdatasets mogelijk om significante thema's, patronen en verborgen verbanden bloot te leggen. Deze techniek is essentieel voor het omzetten van ongestructureerde gegevens in gestructureerde gegevens, die vervolgens kunnen worden gebruikt voor analyse en besluitvorming(**IBM2024**).

2.3.1. Document datatypes

Text mining kan verschillende soorten gegevens omvatten, waaronder:

- **Gestructureerde Gegevens**: Deze gegevens zijn gestandaardiseerd in een tabelvorm, wat ze makkelijker maakt om op te slaan en te verwerken voor analyse en machine learning-algoritmen. Voorbeelden includeren databanken met kolommen en rijen.
- **Ongestructureerde Gegevens**: Deze gegevens hebben geen vooraf gedefinieerd formaat en kunnen tekst uit bronnen zoals sociale media of productrecensies bevatten, evenals rijke media zoals video- en audiobestanden. Aangezien financiële documenten vaak in ongestructureerd formaat bestaan, is text mining essentieel om deze gegevens om te zetten in een bruikbaar formaat.
- **Semi-gestructureerde Gegevens**: Deze gegevens vormen een mix tussen gestructureerde en ongestructureerde formaten. Ze hebben enige organisatie, maar voldoen niet volledig aan de vereisten van een relationele database. Voorbeelden hiervan zijn XML, JSON en Html-bestanden.

Dit onderscheid zijn van groot belang voor het begrijpen van hoe text mining toegepast kan worden over de verschillende datastructuren, dit opent de mogelijkheid om de data de extraheren en belangrijke inzichten te verwerven(**AWS2024**).

2.3.2. Text mining vs. Text analytics

Hoewel text mining en text analytics vaak door elkaar worden gebruikt, kan er een genuanceerd onderscheid tussen de twee gemaakt worden. Bij text mining gaat het meestal om het identificeren van patronen en trends in ongestructureerde gegevens, terwijl text analytics gericht is op het afleiden van kwantitatieve inzichten door gegevens op een gestructureerde manier te analyseren. Deze observaties kunnen vervolgens grafisch worden weergegeven om de ontdekkingen effectief over te brengen aan een breder publiek.(**IBM2024**)

Text mining: vinden van verstopte patronen

Text mining omvat het extraheren van waardevolle informatie en het identificeren van verborgen patronen uit uitgebreide verzamelingen ongeorganiseerde of gedeeltelijk georganiseerde tekstuele gegevens. Text mining is een gespecialiseerde

vorm van datamining die is ontworpen om vooral tekstuele informatie te verwerken. Het belangrijkste doel van text mining is om tekst om te zetten in analyseerbare gegevens om inzichten, trends en patronen te ontdekken die niet direct voor de hand liggen. Deze aanpak omvat een reeks methodologieën, waaronder het ophalen van informatie, natuurlijke taalverwerking (NLP) en machinaal leren. Het primaire doel is het begrijpen en analyseren van uitgebreide tekstdatabases (**gaikwad2014text**).

Voor- en nadelen text mining

Volgens (**Kinter2024**) en (**gaikwad2014text**) zijn er veel voor- en nadelen aan text mining:

- Voordelen:
  - Het corpus van teksten kan worden geanalyseerd met technieken zoals informatie-extractie om de namen van verschillende entiteiten en hun relaties te identificeren.
  - De complexe taak om effectief om te gaan met grote hoeveelheden ongestructureerde gegevens om patronen bloot te leggen, wordt aangepakt door het gebruik van text mining.
  - Bedrijven kunnen een uitgebreid inzicht krijgen in huidige trends en patronen door inzichten te analyseren die zijn verkregen uit vele gegevensbronnen. Deze inzichten helpen bedrijven bij het nemen van weloverwogen zakelijke beslissingen.
- Nadelen
  - Text mining gebruikt vaak een grote hoeveelheid gegevens. Het efficiënt opslaan, beheren en verwerken van deze gegevens vereist daarom een grote hoeveelheid opslagruimte en rekenkracht, wat duur kan zijn.
  - Text mining, gegevensanalyse en patroonherkenning zijn sterk afhankelijk van de kwaliteit van de gegevens. De nauwkeurigheid van de resultaten kan worden beïnvloed door variaties in de gegevenskwaliteit, die worden beïnvloed door de structuur en voorbewerking van de gegevens.

Text analyse: het afleiden van semantische betekenis

Tekstanalyse daarentegen houdt zich meer bezig met het begrijpen en interpreteren van de inhoud van tekst om er informatie van hoge kwaliteit uit af te leiden. In tegenstelling tot text mining, dat zich richt op het ontdekken van nieuwe patronen, is tekstanalyse gericht op het extraheren en interpreteren van bestaande informatie uit tekstgegevens. Dit proces omvat de toepassing van semantische analysetechnieken om de betekenis, context en bedoeling achter de woorden in de tekst te begrijpen (International Journal of Computer Applications, 2014).

Tekstanalyse maakt vaak gebruik van Natural Language Processing (NLP) om de structuur van zinnen te ontleden, entiteiten te identificeren en sentiment te analyseren. Deze technieken zijn cruciaal voor taken zoals sentimentanalyse, waarbij het doel is om de emotionele toon van een tekst te bepalen, of onderwerpmodellering, waarbij het doel is om de belangrijkste thema's te identificeren die in een set documenten worden besproken. Tekstanalyse kan ook meer geavanceerde methoden omvatten, zoals entiteitherkenning, waarbij belangrijke stukken informatie (zoals namen, data en locaties) in een tekst worden geïdentificeerd en geclassificeerd.

conclusie

Terwijl text mining vaak verkennend is, waarbij gezocht wordt naar onbekende patronen, is tekstanalyse gerichter, waarbij de nadruk ligt op het extraheren van specifieke informatie van hoge kwaliteit uit de tekst. In een juridische context kan tekstanalyse bijvoorbeeld worden gebruikt om relevante clausules uit een contract te halen, terwijl text mining kan worden gebruikt om trends in juridische beslissingen in de loop van de tijd te identificeren (International Journal of Computer Applications, 2014).

2.3.3. Text mining technieken

(**Talib2016TextMining**) spreekt over enkele technieken zoals Information Extraction (IE), Information retrieval (IR), en Meerdere NLP technieken die gebruikt worden in data mining, deze zullen hier besproken worden

Information retrieval vs Information extraction

Information retrieval

Informatie retrieval (Information Retrieval, IR) verwijst naar de interactie tussen mens en computer wanneer een gebruiker informatie zoekt die overeenkomt met zijn of haar zoekopdracht in een database of computersysteem. Dit proces omvat het ophalen van relevante inhoud op basis van de behoeften van de gebruiker. Het systeem vergelijkt de zoekopdracht van de gebruiker met een reeks documenten om de meest relevante te identificeren en presenteert deze uiteindelijk in een geprioriteerde lijst. Dit gespecialiseerde vakgebied, zoals beschreven door Krallinger (2024), is essentieel om gebruikers in staat te stellen snel en efficiënt informatie te lokaliseren en extraheren uit uitgebreide en vaak ongestructureerde gegevensbronnen zoals tekstdocumenten, databases of het internet.

De effectiviteit van een IR-systeem wordt gemeten aan de hand van metrieken zoals precisie en recall. Precisie is de verhouding tussen het aantal relevante documenten dat wordt opgehaald en het totale aantal opgehaalde documenten, terwijl recall de verhouding is tussen het aantal relevante documenten dat wordt opgehaald en het totale aantal relevante documenten in de dataset (**Javija2024**). Deze metrics helpen om informatie-overload te verminderen door ervoor te zorgen dat alleen de meest relevante informatie aan de gebruiker wordt gepresenteerd. De

methoden en technieken die worden gebruikt in IR-systemen zijn fundamenteel voor het aandrijven van technologieën zoals zoekmachines, die een snelle en efficiënte informatie ophaling mogelijk maken (**Krallinger2024**).
Enkele IR technieken zijn maar niet gelimiteerd tot (**IBM2024**):

- Tokenizatie: Dit is het proces van het opbreken van text in zinnen en woorden genoemd tokens. Deze zijn dan gebruikt in de modellen voor clustering en documentmatching taken(**IBM2024**).
- Stemming is een tekstvoorbewerkingsmethode die wordt gebruikt in natuurlijke taalverwerking (NLP) om woorden te vereenvoudigen door ze om te zetten naar hun basisvorm. Het doel van stemming is om woorden te stroomlijnen en te normaliseren en zo de efficiëntie van het ophalen van informatie, het categoriseren van teksten en andere natuurlijke taalverwerkingsactiviteiten (NLP) te verbeteren(**SC2024**).

Information extraction
Information Extraction (IE) aims to extract structured information from unstructured documents using techniques like Natural Language Processing (NLP). Unlike Information Retrieval (IR), which retrieves relevant documents, IE focuses on identifying specific data within these texts, making information more accessible and analyseerbare (**Javija2024**). IE systems need to be cost-effective, adaptable, and capable of scaling across domains. In fields like finance, Named-Entity Recognition (NER) is used to extract predefined data types, such as names and dates, from documents, facilitating efficient data management (Gupta2020). Automated learning in IE reduces errors and dependency on manual supervision, making the process more efficient and contextually valuable. The increasing volume of unstructured data, particularly online, emphasizes the importance of effective IE systems (**Javija2024**).
Enkele IE technieken zijn maar niet gelimiteerd tot (**IBM2024**):

- Feature selection en Feature extraction
  - Feature selection: is een essentiële stap bij het verwerken van gegevens met een groot aantal dimensies. Het gaat om het kiezen van een kleinere set belangrijke kenmerken uit de originele set om de efficiëntie en nauwkeurigheid van het leren te verbeteren. Door overbodige en inconsequente kenmerken te elimineren, wordt de omvang van de gegevensverwerking verkleind, wordt de tijd die nodig is voor het leren verminderd en worden de resultaten gestroomlijnd. Eigenschapsselectie is een proces dat de belangrijkste oorspronkelijke kenmerken behoudt, in tegenstelling tot kenmerkextractie waarbij gegevens worden veranderd in kenmerken die goed zijn in het herkennen van patronen. Eigenschapsselec-

tie is cruciaal voor het verminderen van de dimensionaliteit van gegevens. Technieken voor kenmerkselectie omvatten een reeks benaderingen, zoals supervised, unsupervised en semi-supervised modellen. Deze methoden worden geclassificeerd op basis van hun associatie met leermethoden (filter, wrapper, inbeddingsmodellen) en andere criteria. Eigenschapsselectie is een veelgebruikte techniek in gebieden zoals beeldherkenning en tekst mining. Het verbetert de prestaties van modellen voor machinaal leren door een evenwicht te bereiken tussen hoge nauwkeurigheid en lage rekenvereisten (**CAI201870**).

- Feature extraction: is een essentiële stap in machinaal leren, omdat het uitgebreide invoergegevens omzet in een beter hanteerbare en lager-dimensionale kenmerkenset. Deze procedure vereenvoudigt de gegevens door de complexiteit ervan te verminderen, terwijl belangrijke informatie toch behouden blijft. Het is vooral nuttig bij taken zoals categorisatie. Kenmerkextractietechnieken transformeren de initiële kenmerkruimte in een gecondenseerde, alternatieve ruimte door een gereduceerde, representatieve verzameling kenmerken te behouden in plaats van ze weg te gooien. Principale Componenten Analyse (PCA) en Bag of Words zijn vaak gebruikte technieken. PCA vermindert bijvoorbeeld de dimensionaliteit van gegevens door de oorspronkelijke variabelen om te zetten in ongecorrigeerde componenten. Dit proces verbetert de rekenefficiëntie en verhoogt de nauwkeurigheid van modellen voor machinaal leren (**Mustazzihim**).
- Verschil? FS behoud de originele features terwijl FE nieuwe maakt.

- Named Entity Recognition (NER) is a core job in Natural Language Processing (NLP) that aims to recognise and categorise entities, such as individuals, organisations, and places, within a given text. NER, or Named Entity Recognition, is extensively utilised in a variety of applications, spanning from information retrieval to automated customer care.

  Recent research emphasises that although NER models have attained remarkable performance on typical datasets, frequently exhibiting high F-scores, this measure alone does not offer a thorough comprehension of their efficacy. For instance, cutting-edge NER models typically exhibit F-scores over 90% on datasets such as OntoNotes. Nevertheless, this solitary metric may obscure variations in performance across various categories of entities, kinds of language, and unfamiliar data (**vajjala2022reallyknowstateart**).

conclusie

Information Retrieval (IR) en Information Extraction (IE) zijn twee technologieën die informatie toegankelijk maken via verschillende methodologieën.IR richt zich op het ophalen van relevante documenten, terwijl IE specifieke, gestructureerde in-

formatie extraheert voor nauwkeurige gegevensanalyse.IR is essentieel voor grote datasets en zoekmachines, terwijl IE cruciaal is voor het extraheren van bruikbare inzichten. De kracht van IR ligt in het beheren en ophalen van informatie uit ongestructureerde bronnen, waardoor het onmisbaar is voor grote databases. IE is van vitaal belang voor datamining, kennisbeheer en geautomatiseerde processen. Naarmate het datavolume toeneemt, zal de wisselwerking tussen IR en IE steeds belangrijker worden. Het begrijpen en benutten van beide technologieën zal cruciaal zijn voor het optimaliseren van informatieverwerkingssystemen en om ervoor te zorgen dat gebruikers snel en accuraat de benodigde informatie kunnen verkrijgen.

Aspect	**Information Retrieval**	**Information Extraction**
Focus	Document Retrieval	Feature Retrieval
Output	Return set of relevant documents	Return facts out of documents
Goal	The goal is to find documents that are relevant to the user's information need	The goal is to extract prespecified features from documents or display information.
Nature of Information	Real information is buried inside documents	Extract information from within the documents
Result Format	The long listing of documents	Aggregate over the entire set
Application	Used in many search engines – Google is the best IR system for the web.	Used in database systems to enter extracted features automatically.
Methodology	Typically uses a bag of words model of the source text.	Typically based on some form of semantic analysis of the source text.
Theoretical Basis	Mostly use the theory of information, probability, and statistics.	Emerged from research into rule-based systems.

Tabel 2.1: Comparison of Information Retrieval and Information Extraction

NLP

Summarization

Een andere kritische NLP techniek is tekstsamenvatting, waarbij een beknopte weergave van originele tekstdocumenten wordt gegenereerd. Dit proces omvat voorbewerkingsstappen zoals tokeniseren, stopwoorden verwijderen en stemmen,

gevolgd door het creëren van lexiconlijsten tijdens de verwerkingsfase. Historisch gezien was het samenvatten van tekst gebaseerd op woordfrequentie, maar moderne methoden maken gebruik van geavanceerde text mining technieken om de relevantie en nauwkeurigheid van de resultaten te verbeteren. Kenmerken zoals zinslengte, thematische woorden en vaste zinnen worden gebruikt om belangrijke informatie te extraheren en deze technieken kunnen op meerdere documenten tegelijk worden toegepast(**Talib2016TextMining**).

Part of speech tagging

Part-of-speech (POS) tagging is een essentiële activiteit in natuurlijke taalverwerking (NLP) waarbij een grammaticale classificatie, zoals zelfstandig naamwoord, werkwoord of bijvoeglijk naamwoord, wordt toegewezen aan elk woord in een zin. Tagging vergemakkelijkt computationeel begrip van de syntactische organisatie van tekst, een kritisch onderdeel voor veel toepassingen van natuurlijke taalverwerking (NLP) (Martinez,2012).Ondanks de uitdagingen zoals tweeslachtige woorden bereiken moderne POS taggers hoge nauwkeurigheidspercentages (rond 96-97%) en worden ze veel gebruik bij het ophalen van informatie, tekstanalyse en andere NLP-taken(**Martinez2024**).

Text categorization

Tekstclassificatie is een methodische procedure die bestaat uit vier essentiële stappen: kenmerken extraheren, de dimensionaliteit verminderen, een classificator selecteren en de resultaten evalueren. Eerst wordt tekst omgezet in een numeriek formaat door middel van kenmerkextractie, zoalswoordfrequentieofWord2Vec.Dimensionaliteitsreduc gebruikt om de gegevens te vereenvoudigen en cruciale informatie te behouden. Dit wordt bereikt door technieken zoals principale componentenanalyse(PCA) of lineaire discriminantanalyse (LDA) toe te passen. De selectie van een classificeer is essentieel, omdat deep learning-methoden vaak conventionele machinelearning-algoritmen overtreffen in termen van nauwkeurigheid. Uiteindelijk wordt de doeltreffendheid van het model beoordeeld door de prestaties te meten met behulp van metrieken zoals de Matthews correlatiecoëfficiënt (MCC), oppervlakte onder de ROC-curve (AUC) en nauwkeurigheid. Van deze maatstaven wordt nauwkeurigheid beschouwd als de meest directe en eenvoudige manier om de prestatie van het model te evalueren. Gupta et al. (2020) ontdekten dat supportvectormachines (SVM) beter presteerden dan andere benaderingen zoals Naive Bayes (NB),k-nearest neighbour (KNN), beslisbomen en regressie in termen van nauwkeurigheid, recall en F1-maatstaven.

Sentiment analysis

Natural Language Processing (NLP) omvat verschillende technieken om onnauwkeurig en dubbelzinnig taalgebruik om te zetten in nauwkeurige en ondubbelzin-

nige berichten, met toepassingen in sectoren als financiën, e-commerce en sociale media. Een belangrijke techniek binnen NLP is Sentimentanalyse (SA), ook wel bekend als opinion mining, waarbij onderliggende meningen uit tekstgegevens worden gehaald. SA is vooral nuttig voor taken als emotieherkenning en polariteitsdetectie, met toepassingen variërend van voorspelling van de aandelenmarkt tot analyse van feedback van klanten (**Gupta2020**).
SA kan worden benaderd met lexicon gebaseerde methoden, die vertrouwen op tools zoals SentiWordNet voor woord-naar-sentiment mappen, of machine learning (ML) technieken die tekst classificeren met behulp van algoritmes zoals Naïve Bayes (NB) en support vectormachines (SVM's). Hoewel ML-benaderingen geen kostbare woordenboeken vereisen, vereisen ze wel domein specifieke datasets, wat een beperking kan zijn. Bovendien zijn deep learning-methoden onlangs gecombineerd met traditionele ML-technieken om de nauwkeurigheid en betrouwbaarheid van SA te verbeteren, met name in financiële voorspellingen (**Gupta2020**).

Testing

2.4. Technieken en Tools

2.4.1. SpaCy vs NLTK

Tabel geeft een vergelijkende analyse van SpaCy en NLTK op basis van belangrijke functies die relevant zijn voor tekstsamenvatting. (**amade2024automatic**)
Prestatiewaarden (Precisie, Recall, F-Score): SpaCy toont hogere precisie (0.72 vs. 0.51) en F-Score (0.69 vs. 0.58) in vergelijking met NLTK, wat wijst op superieure nauwkeurigheid bij het genereren van samenvattingen.
Snelheid van Tokenizatie en Tagging: SpaCy is aanzienlijk sneller dan NLTK in zowel tokeniseren als taggen. SpaCy kan bijvoorbeeld tekst tokenen in 0,2 milliseconden vergeleken met de 4 milliseconden van NLTK, waardoor het geschikter is voor toepassingen die real-time verwerking vereisen. **Ondersteuning voor Geavanceerde NLP-functies:** SpaCy ondersteunt geavanceerde functies zoals topic modellering, vectorisatie en TF-IDF (Term Frequency-Inverse Document Frequency), die niet standaard beschikbaar zijn in NLTK. Dit maakt SpaCy een meer uitgebreide tool voor taken die diep semantisch begrip en machine learning integratie vereisen.
Gebruiksvriendelijkheid: SpaCy is gebruiksvriendelijker met ingebouwde mogelijkheden, wat de behoefte aan maatwerkprogrammering vermindert, in tegenstelling tot NLTK, dat meer tijd en inspanning vergt.

subsubsection

Conclusie Samenvattend, SpaCy is een krachtigere en efficiëntere tool voor tekstsamenvatting vanwege zijn hogere precisie, snelheid en ondersteuning voor geavanceerde NLP-functies. NLTK, hoewel veelzijdig, is beter geschikt voor eenvoudigere taken of projecten die meer aanpassing vereisen. De keuze tussen deze tools hangt

Functie	NLTK	SpaCy
Precisie	0.51	0.72
Recall	0.65	0.65
F-Score	0.58	0.69
Tokenisatiesnelheid	4 ms	0.2 ms
Taggingsnelheid	443 ms	1 ms
Ondersteuning voor Classificatie	Ja	Ja
Onderwerpmodellering	Nee	Ja
Vectorisatie	Nee	Ja
Parsing	Ja	Ja
TF-IDF Implementatie	Nee	Ja
Programmeerparadigma	Procedureel	Objectgeoriënteerd
Gebruiksvriendelijkheid	Vereist meer aanpassing en tijd	Meer geautomatiseerd en gebruiksvriendelijk
Ondersteunde Taalmodellen	Basis Tokenizatie en parsing	Geavanceerde modellen met voorgetrainde vectors
Grootte en Afhankelijkheden van de Bibliotheek	Lichtgewicht, minimale afhankelijkheden	Zwaarder, meer afhankelijkheden door geavanceerde functies

Tabel 2.2: Vergelijkende Analyse van SpaCy en NLTK

af van de specifieke eisen van het project, waaronder de complexiteit van de taak, de benodigde functies en de beschikbare middelen.

2.4.2. Database Management Systemen (DBMS)

In deze sectie gaan wij bekijken welke databank gebruikt zal worden na het structuren en standaardiseren van de 13f meldingen. Hier zal besproken worden of er SQL of nosql gebruikt zal worden vervolgens zal er een specifieke databank gekozen worden

SQL vs. NOSQL

Bij het kiezen tussen SQL- en Nosql-databases is het belangrijk om de onderliggende architectuur en toepassingsmogelijkheden te begrijpen. SQL-databases zijn ontworpen voor het organiseren van gestructureerde data, waardoor ze ideaal zijn voor online transaction processing (OLTP). Ze presteren uitstekend in situaties waarin

complexe query's, consistentie en relationeel databeheer vereist zijn. Nosql-databases daarentegen ondersteunen horizontale schaalbaarheid en zijn geoptimaliseerd voor het verwerken van grote hoeveelheden ongestructureerde data, wat hen geschikt maakt voor big data-analyse. De keuze tussen beide hangt grotendeels af van de specifieke behoeften van de organisatie, zoals de focus op datastructuur of schaalbaarheid (**khan2023performance**).
In dit onderzoek is gekozen voor een SQL-database. Deze keuze is gebaseerd op de noodzaak om gestructureerde data uit de 13F-meldingen te beheren, waarbij consistente gegevensintegriteit en de mogelijkheid om complexe query's uit te voeren cruciaal zijn. SQL-databases bieden de benodigde functionaliteiten voor het beheer van relationele gegevens en het uitvoeren van geavanceerde analyses, wat essentieel is voor het succes van dit project (**khan2023performance**).

SQL-databank
Op basis van de gedetailleerde analyse in het GeeksforGeeks-artikel werd PostgreSQL gekozen voor ons proefschrift vanwege de geavanceerde functies, robuuste gegevensintegriteit en uitbreidbare architectuur. In tegenstelling tot andere SQL-databases, blinkt PostgreSQL uit in het verwerken van complexe datamanipulatie, het bieden van sterke ACID compliance en het ondersteunen van aangepaste datatypes en functies. Dit maakt PostgreSQL bijzonder geschikt voor bedrijfstoepassingen en datawarehousing waar schaalbaarheid en geavanceerd databeheer cruciaal zijn. Hoewel PostgreSQL een steilere leercurve heeft dan sommige alternatieven, maken de uitgebreide functie set en betrouwbaarheid het een optimale keuze om aan de complexe eisen van ons project te voldoen.

Review
Dus voor dit onderzoek hebben we voor de data mining gekozen voor SpaCy en voor de DB PostgreSQL

2.5. Uitdagingen en beperkingen

2.5.1. Complexiteit van financiële Tekst

2.5.2. Gegevenskwaliteit en Validatie

2.5.3. Databaseprestaties

Uitdaging: Hoewel PostgreSQL goed presteert bij grote hoeveelheden gestructureerde data, kan het moeilijk zijn om de prestaties te optimaliseren naarmate de hoeveelheid data en het aantal gelijktijdige gebruikers toeneemt.
Beperking: Bij zeer grote datasets of een hoge mate van gelijktijdige toegang kunnen er prestatieproblemen optreden. Het kan nodig zijn om uitgebreide optimalisaties en schaalstrategieën te implementeren, zoals partitionering of het gebruik van read replicas.

2.6. Leemtes in huidig onderzoek

2.6.1. Onbehandelde kwesties

Training van Eigen Large Language Models (LLMs) Het trainen van een eigen LLM voor financiële toepassingen vereist veel tijd en middelen. Het model moet worden getraind op uitgebreide financiële datasets voor nauwkeurige resultaten. Dit kan uw infrastructuur belasten en vereist expertise in machine learning en datawetenschap, met mogelijke problemen op het gebied van data-integriteit en privacy. Overweeg het gebruik van bestaande financiële NLP-modellen die al getraind zijn en efficiënt kunnen worden aangepast aan uw behoeften.
Beveiliging en Privacy Het beschermen van gevoelige financiële gegevens tegen ongeautoriseerde toegang en datalekken is complex en vereist naleving van privacywetgeving. Onvoldoende beveiliging kan leiden tot datalekken, verlies van vertrouwen en juridische problemen. Implementeer encryptie, toegangscontrole en regelmatige beveiligingsaudits om gegevens te beschermen. Zorg ervoor dat uw systemen voldoen aan relevante regelgeving en best practices voor gegevensbeveiliging.
Schaalbaarheid en Prestaties Groeiende hoeveelheden gegevens kunnen leiden tot prestatieproblemen bij opslag en analyse, wat complexe oplossingen vereist voor snelle toegang. Slechte prestaties kunnen vertragingen veroorzaken in rapportage en analyse, wat de besluitvorming en efficiëntie beïnvloedt. Gebruik schaalbare databases en technieken zoals gegevenspartitionering en caching. Monitor en optimaliseer regelmatig de prestaties om problemen te voorkomen.

2.6.2. Verbeteringsmogelijkheden

2.7. conclusie

2.7.1. Samenvatting van Bevindingen

2.7.2. Implicaties van het onderzoek

3

Methodologie

Dit hoofdstuk geeft een overzicht van de methodologie die is gebruikt om dit onderzoek uit te voeren en de Proof of Concept (POC) te creëren. De tekst biedt een uitgebreide analyse van het belang van elke fase van het onderzoek en licht de redenering achter de gekozen methodologieën en benaderingen toe. Dit hoofdstuk maakt duidelijk hoe de gekozen benaderingen helpen om de onderzoeksdoelen te bereiken door een goed georganiseerd overzicht te bieden. Het belang van elke fase wordt benadrukt, waardoor inzicht wordt verkregen in de achterliggende gedachte van de beslissingen die tijdens het onderzoeksproces zijn genomen.

3.1. Literatuur studie

De eerste fase van dit onderzoek bestond uit een uitgebreid onderzoek van bestaande literatuur. Het doel van deze fase was om een grondig begrip te krijgen van de concepten en technologieën die gebruikt zouden worden bij de implementatie van de Proof of Concept (POC). De bovengenoemde stap omvatte een uitgebreide analyse van verschillende publicaties, papers, blogs en handleidingen om relevante toepassingen en benaderingen te ontdekken. De belangrijkste onderwerpen die in deze fase werden onderzocht waren Text Mining, Natural Language Processing (NLP) en Database Management Systemen (DBMS).

3.2. Requirements analyse

3.3. Dataset creation

Tijdens deze fase hebben we een dataset gegenereerd met de 13F-dossiers als basis, die de basis vormde voor de constructie van het Proof of Concept (POC). De informatie werd zorgvuldig samengesteld door pertinente financiële gegevens uit de 13F-papieren te halen, waarbij gegarandeerd werd dat de informatie geordend

en geformatteerd werd op een manier die geschikt is voor latere analyse en verwerking binnen het POC-kader.

3.4. POC

In het volgende deel van ons onderzoek willen we een Proof of Concept (POC) uitvoeren met SpaCy, een zeer gewaardeerd Python framework voor natuurlijke taalverwerking (NLP). SpaCy werd geselecteerd vanwege zijn sterke vaardigheid in het beheren van uitgebreide tekstdatasets, wat cruciaal is voor het analyseren van ingewikkelde financiële informatie, zoals die in 13F filings.
Het doel van de proof of concept (POC) is het onderzoeken en verifiëren van de doeltreffendheid van SpaCy bij het uitvoeren van essentiële NLP (natural language processing) activiteiten, zoals het extraheren van tekst, het herkennen van named entities (NER) en het ophalen van informatie. Het uitvoeren van deze taken is cruciaal voor de nauwkeurige identificatie en extractie van belangrijke entiteiten, zoals bedrijfsnamen en financiële statistieken, uit het ongeorganiseerde materiaal in 13F-papers.
Een cruciaal onderdeel van deze proof of concept (POC) is het aanpassen van SpaCy's natuurlijke taalverwerking (NLP) aan de specifieke eisen van het project. Een van deze benaderingen is het trainen van gespecialiseerde NER-modellen met behulp van domeinspecifieke gegevens. Dit kan helpen om de nauwkeurigheid van het herkennen van financiële woorden en entiteiten te verhogen, waardoor de algehele betrouwbaarheid van het systeem toeneemt.
Door de implementatie van deze proof of concept (POC) willen we de haalbaarheid van het gebruik van SpaCy voor dit project aantonen en een basis leggen voor de uitgebreide implementatie van de natuurlijke taalverwerkingsoplossingen (NLP) die nodig zijn voor het verwerken en onderzoeken van financiële gegevens uit 13F filings.

3.5. Database

PostgreSQL is gekozen als databasebeheeroplossing voor dit project om te voldoen aan de vereisten voor gegevensbeheer. PostgreSQL is een vrij beschikbaar databasesysteem dat de eigenschappen van objectgeoriënteerde en relationele databases combineert. Het is zeer betrouwbaar, kan grote hoeveelheden gegevens aan en heeft uitstekende mogelijkheden voor het uitvoeren van gecompliceerde queries. Deze kwaliteiten maken het een uitstekende optie voor het beheren van de gedetailleerde financiële informatie uit 13F deponeringen.
De keuze voor PostgreSQL werd beïnvloed door drie belangrijke factoren:
PostgreSQL garandeert de integriteit van gegevens door de ACID (Atomicity, Consistency, Isolation, Durability) principes volledig te ondersteunen. Precisie en consistentie zijn van het grootste belang bij het verwerken van gevoelige financiële

gegevens.
PostgreSQL heeft geavanceerde mogelijkheden, waaronder ondersteuning voor JSON-gegevenstypen, full-text zoeken en aangepaste indexering. Deze functies zijn met name voordelig voor het effectief verwerken van semigestructureerde gegevens die kunnen voortkomen uit natuurlijke taalverwerkingstaken (NLP).
Het flexibele ontwerp van PostgreSQL maakt het mogelijk om de database aan te passen aan de unieke eisen van ons project, met name wat betreft het opslaan en opvragen van financiële informatie. Dit wordt bereikt door de mogelijkheid om nieuwe functies, operatoren en datatypes te bouwen.
Schaalbaarheid is een belangrijke factor bij het overwegen van PostgreSQL's vermogen om enorme datasets goed te beheren, vooral gezien de omvang en complexiteit van de gegevens die geproduceerd worden tijdens het verwerken van 13F aanvragen. PostgreSQL heeft de mogelijkheid om zowel horizontaal als verticaal uit te breiden, zodat het effectief kan voldoen aan de toenemende eisen van het project naarmate er meer gegevens worden toegevoegd.

3.6. Analyse van de resultaten

4

Conclusie

A

Onderzoeksvoorstel

A.1. Inleiding

Waarover zal je bachelorproef gaan? Introduceer het thema en zorg dat volgende zaken zeker duidelijk aanwezig zijn:

- kaderen thema
- de doelgroep
- de probleemstelling en (centrale) onderzoeksvraag
- de onderzoeksdoelstelling

Denk er aan: een typische bachelorproef is *toegepast onderzoek*, wat betekent dat je start vanuit een concrete probleemsituatie in bedrijfscontext, een **casus**. Het is belangrijk om je onderwerp goed af te bakenen: je gaat voor die *ene specifieke probleemsituatie* op zoek naar een goede oplossing, op basis van de huidige kennis in het vakgebied.
De doelgroep moet ook concreet en duidelijk zijn, dus geen algemene of vaag gedefinieerde groepen zoals *bedrijven*, *developers*, *Vlamingen*, enz. Je richt je in elk geval op it-professionals, een bachelorproef is geen populariserende tekst. Eén specifiek bedrijf (die te maken hebben met een concrete probleemsituatie) is dus beter dan *bedrijven* in het algemeen.
Formuleer duidelijk de onderzoeksvraag! De begeleiders lezen nog steeds te veel voorstellen waarin we geen onderzoeksvraag terugvinden.
Schrijf ook iets over de doelstelling. Wat zie je als het concrete eindresultaat van je onderzoek, naast de uitgeschreven scriptie? Is het een proof-of-concept, een rapport met aanbevelingen, ...Met welk eindresultaat kan je je bachelorproef als een succes beschouwen?

A.2. Literatuurstudie

Hier beschrijf je de *state-of-the-art* rondom je gekozen onderzoeksdomein, d.w.z. een inleidende, doorlopende tekst over het onderzoeksdomein van je bachelorproef. Je steunt daarbij heel sterk op de professionele *vakliteratuur*, en niet zozeer op populariserende teksten voor een breed publiek. Wat is de huidige stand van zaken in dit domein, en wat zijn nog eventuele open vragen (die misschien de aanleiding waren tot je onderzoeksvraag!)?
Je mag de titel van deze sectie ook aanpassen (literatuurstudie, stand van zaken, enz.). Zijn er al gelijkaardige onderzoeken gevoerd? Wat concluderen ze? Wat is het verschil met jouw onderzoek?
Verwijs bij elke introductie van een term of bewering over het domein naar de vakliteratuur, bijvoorbeeld (**Hykes2013**)! Denk zeker goed na welke werken je refereert en waarom.
Draag zorg voor correcte literatuurverwijzingen! Een bronvermelding hoort thuis *binnen* de zin waar je je op die bron baseert, dus niet er buiten! Maak meteen een verwijzing als je gebruik maakt van een bron. Doe dit dus *niet* aan het einde van een lange paragraaf. Baseer nooit teveel aansluitende tekst op eenzelfde bron.
Als je informatie over bronnen verzamelt in JabRef, zorg er dan voor dat alle nodige info aanwezig is om de bron terug te vinden (zoals uitvoerig besproken in de lessen Research Methods).
Je mag deze sectie nog verder onderverdelen in subsecties als dit de structuur van de tekst kan verduidelijken.

A.3. Methodologie

Hier beschrijf je hoe je van plan bent het onderzoek te voeren. Welke onderzoekstechniek ga je toepassen om elk van je onderzoeksvragen te beantwoorden? Gebruik je hiervoor literatuurstudie, interviews met belanghebbenden (bv. voor requirementsanalyse), experimenten, simulaties, vergelijkende studie, risico-analyse, PoC, ...?
Valt je onderwerp onder één van de typische soorten bachelorproeven die besproken zijn in de lessen Research Methods (bv. vergelijkende studie of risico-analyse)? Zorg er dan ook voor dat we duidelijk de verschillende stappen terug vinden die we verwachten in dit soort onderzoek!
Vermijd onderzoekstechnieken die geen objectieve, meetbare resultaten kunnen opleveren. Enquêtes, bijvoorbeeld, zijn voor een bachelorproef informatica meestal **niet geschikt**. De antwoorden zijn eerder meningen dan feiten en in de praktijk blijkt het ook bijzonder moeilijk om voldoende respondenten te vinden. Studenten die een enquête willen voeren, hebben meestal ook geen goede definitie van de populatie, waardoor ook niet kan aangetoond worden dat eventuele resultaten representatief zijn.
Uit dit onderdeel moet duidelijk naar voor komen dat je bachelorproef ook tech-

nisch voldoende diepgang zal bevatten. Het zou niet kloppen als een bachelorproef informatica ook door bv. een student marketing zou kunnen uitgevoerd worden.
Je beschrijft ook al welke tools (hardware, software, diensten, ...) je denkt hiervoor te gebruiken of te ontwikkelen.
Probeer ook een tijdschatting te maken. Hoe lang zal je met elke fase van je onderzoek bezig zijn en wat zijn de concrete *deliverables* in elke fase?

A.4. Verwacht resultaat, conclusie

Hier beschrijf je welke resultaten je verwacht. Als je metingen en simulaties uitvoert, kan je hier al mock-ups maken van de grafieken samen met de verwachte conclusies. Benoem zeker al je assen en de onderdelen van de grafiek die je gaat gebruiken. Dit zorgt ervoor dat je concreet weet welk soort data je moet verzamelen en hoe je die moet meten.
Wat heeft de doelgroep van je onderzoek aan het resultaat? Op welke manier zorgt jouw bachelorproef voor een meerwaarde?
Hier beschrijf je wat je verwacht uit je onderzoek, met de motivatie waarom. Het is **niet** erg indien uit je onderzoek andere resultaten en conclusies vloeien dan dat je hier beschrijft: het is dan juist interessant om te onderzoeken waarom jouw hypothesen niet overeenkomen met de resultaten.

B

Bijlagen