Onderzoeksvoorstel

Vul hier de voorgestelde titel van je onderzoek in.

Bachelorproef, 2022-2023

Thomas Vanderveken

E-mail: thomas.vanderveken@student.hogent.be
Co-promotor: Dhr. L. Smits (lieven.smits@synalco.be)

Samenvatting

Deze bachelorproef onderzoekt hoe AI-technologieën, zoals Natural Language Processing (NLP) en Machine Learning (ML), gebruikt kunnen worden om 13F-meldingen van de SEC van voor het jaar 2013 te standaardiseren. Dit kan historisch financieel onderzoek en investeringsanalyse vergemakkelijken, omdat deze documenten momenteel allemaal handmatig moeten worden bekeken. Het doel is om efficiënte en nauwkeurige data-extractie uit deze meldingen te realiseren, wat essentieel is voor het verkrijgen van inzichten in historische beleggingstrends. Door middel van een literatuurstudie zal onderzocht worden welke NLP-technieken en ML-modellen het meest effectief zijn voor deze taak. Het uiteindelijke doel is om een proof-of-concept te ontwikkelen die deze 13F-meldingen verwerkt, de benodigde gegevens extraheert en deze opslaat in een databank voor gemakkelijke toegang tot de informatie.

Keuzerichting: AI& Data Engineering
Sleutelwoorden: 13F, TextMining, ML

Inhoudsopgave

1. Inleiding

1.1. Achtergrond en Context

13F-meldingen, bij de SEC zijn ingediend, bevatten essentiële informatie over de beleggingsportefeuilles van institutionele investeerders en zijn van cruciaal belang voor financieel onderzoek en investeringsanalyse. Maar voorafgaand aan 2013 vertonen 13F-rapporten vaak inconsistenties in formaat en structuur, waardoor handmatige verwerking extreem tijdrovend en foutgevoelig is.

AI-technologieën zoals NLP en ML kunnen helpen deze oudere documenten te standaardiseren en vervolgens te integreren in een gestructureerde databank. Dit zou de efficiëntie van gegevensverwerking verbeteren en de toegankelijkheid van historische financiële data vergroten. Een proof-of-concept applicatie die deze AI-technieken toepast, zal niet alleen de analyse van historische beleggingstrends vergemakkelijken, maar ook het ontwikkelen van voorspellende modellen eenvoudiger maken.

1.2. 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.

1.3. Hoofonderzoeksvraag

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

de historische gegevens efficiënter kunnen worden geanalyseerd en vergeleken?

1.4. Deelonderzoeksvragen

1. Wat zijn de potentiële voordelen en beperkingen van het gebruik van AI-technologieën voor dit doel vergeleken met traditionele methoden?
2. Wat zijn de belangrijkste uitdagingen bij het standaardiseren van de verschillende formaten en structuren van 13F-meldingen?
3. Hoe kan de ontwikkelde proof-of-concept worden gevalideerd en geëvalueerd op basis van nauwkeurigheid, efficiëntie en bruikbaarheid?

1.5. 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.

2. Literatuurstudie

In dit gedeelte van het voorstel worden verschillende componenten behandeld. Allereerst zal de focus liggen op de rol van de SEC en de aard van 13F-meldingen. Vervolgens zullen we ingaan op text mining en de diverse technieken die daarbij worden toegepast.

2.1. SEC en 13F

2.1.1. Definitie en doel

13F-meldingen zijn wettelijke rapporten die de Amerikaanse Securities and Exchange Commission (SEC) vereist onder Sectie 13(f) van de Securities Exchange Act van 1934. Ze dienen om de portefeuilles van institutionele beleggingsbeheerders te rapporteren Securities en Commission (2023). Het belangrijkste doel van deze meldingen is om transparantie te waarborgen over de beleggingsactiviteiten van grote institutionele beleggers zoals beleggingsfondsen en pensioenfondsen. Dit helpt zowel het publiek als regelgevende instanties om toezicht te houden op de beleggingsposities van deze instellingen.

2.1.2. Belangrijke kenmerken

Vereisten voor rapportage: Institutionele beleggers met een beheerd vermogen van minimaal USD 100 miljoen moeten elk kwartaal een 13F-melding indienen. Deze rapporten bevatten gedetailleerde informatie over hun aandelenportefeuille, zoals de naam van het aandeel, het CUSIP-nummer, het aantal gehouden aandelen en de marktwaarde ervan.

Omvang van de informatie: De rapportages zijn voornamelijk gericht op aandelenbezit. Andere activa, zoals obligaties, derivaten en private equity, worden niet opgenomen in de meldingen Securities en Commission (2023). Elk kwartaalrapport biedt een overzicht van de aandelenportefeuille aan het einde van de rapportageperiode, wat waardevolle inzichten geeft in de investeringsstrategieën en methoden van de instelling.

Opmaak en toegankelijkheid: De 13F-meldingen voor 2013 en eerder hebben een variërende opmaak, wat het extractieproces van gegevens bemoeilijkt. Dit maakt het lastig om consistente en betrouwbare gegevens te verkrijgen uit deze oudere rapporten.

Deze meldingen dragen bij aan de transparantie en helpen bij het begrijpen van de beleggingsbenaderingen van grote institutionele beleggers.

2.2. Textmining

Text mining, ook wel tekst datamining genoemd, is het proces waarbij ongestructureerde tekst wordt omgezet in een gestructureerd formaat om patronen te ontdekken en nieuwe inzichten te verwerven (IBM, 2024). Deze techniek maakt het mogelijk om uit uitgebreide tekst datasets significante thema's, patronen en verborgen verbanden te extraheren, wat cruciaal is voor analyse en besluitvorming.

2.3. Documentdatatypes

Text mining kan verschillende soorten gestructureerde gegevens omvatten, zoals:

1. **Gestructureerde Gegevens:** Georganiseerd in tabelvorm, wat verwerking en analyse vergemakkelijkt, zoals in databanken met kolommen en rijen.
2. **Ongestructureerde Gegevens:** Tekst zonder vooraf gedefinieerd formaat, zoals sociale media of productrecensies, en rijke media zoals video- en audiobestanden. Financiële documenten vallen vaak onder deze categorie, waardoor text mining essentieel is om deze gegevens bruikbaar te maken.
3. **Semi-gestructureerde Gegevens:** Een mix van gestructureerde en ongestructureerde formaten, zoals XML, JSON en HTML-bestanden, die enige organisatie hebben maar niet volledig voldoen aan relationele databasevereisten (AWS, 2024).

Het begrijpen van deze datatypes is cruciaal voor het toepassen van text mining op verschil-

lende datastructuren, wat mogelijkheden opent voor het extraheren van belangrijke inzichten.

2.4. Text Mining versus Text Analytics

Hoewel text mining en text analytics vaak door elkaar worden gebruikt, zijn er nuances in hun toepassingen. Text mining richt zich op het ontdekken van patronen en trends in ongestructureerde gegevens, terwijl text analytics zich richt op het afleiden van kwantitatieve inzichten door gestructureerde gegevens te analyseren (IBM, 2024). Text mining omvat methoden zoals informatie-extractie, natuurlijke taalverwerking (NLP) en machinaal leren om verborgen patronen te ontdekken in grote hoeveelheden tekstuele gegevens (Gaikwad e.a., 2014).

2.5. Voor- en Nadelen van Text Mining

Voordelen:

1. Analyse van grote tekst corpora om entiteiten en hun relaties te identificeren.
2. Omgaan met ongestructureerde gegevens om patronen te ontdekken.
3. Inzichten uit verschillende gegevensbronnen voor weloverwogen zakelijke beslissingen.

Nadelen:

1. Vereist aanzienlijke opslagruimte en rekenkracht.
2. Resultaten zijn afhankelijk van de gegevenskwaliteit, beïnvloed door structuur en voorbewerking (Gaikwad e.a., 2014; Kinter, 2024).

2.6. Tekstanalyse versus Text Mining

Tekstanalyse richt zich op het extraheren en interpreteren van specifieke informatie uit tekstgegevens, met behulp van semantische analysetechnieken en NLP voor taken zoals sentimentanalyse en onderwerpmodellering (Gaikwad e.a., 2014). Dit kan bijvoorbeeld worden gebruikt om relevante clausules uit contracten te halen, terwijl text mining wordt gebruikt om trends in juridische beslissingen te identificeren.

In conclusie, text mining en tekstanalyse dienen verschillende doeleinden: text mining zoekt naar onbekende patronen, terwijl tekstanalyse zich richt op het extraheren van bestaande informatie van hoge kwaliteit.

3. Methodologie

Dit onderzoek richt zich op het ontwikkelen van een proof-of-concept applicatie die AI- technologieën, zoals Natural Language Processing (NLP) en Machine Learning (ML), gebruikt om 13F-meldingen van voor 2013 te standaardiseren en te integreren in een relationele databank. De methodologie omvat vier hoofdfasen: literatuurstudie, systeemontwikkeling, evaluatie, en implementatie.

In de eerste fase zal de literatuurstudie worden voorbereid, deze zal zich focussen op het analyseren van bestaande technieken en benaderingen te analyseren. Dit zal bestaan uit het verkennen van relevante NLP-technieken zoals Named Entity Recognition (NER), tekstclassificatie en tokenisatie, die nuttig kunnen zijn voor het extraheren van de nodige gegevens. Alsook zal er een analyse gedaan worden naar al bestaande modellen en bibliotheken zoals BERT, GPT en Spacy.

Op basis van de bevindingen uit de eerste fase zal er een proof-of-concept systeem ontwikkeld met de volgende stappen:

1. **Data Voorbereiding:** Verzamelen en voorbereiden van een dataset van 13F- meldingen van voor 2013. Dit kan bestaan uit het downloaden van historische rapporten en het opschonen van gegevens om consistentie en kwaliteit te waarborgen.
2. **NLP- en ML-implementatie:** Het toepassen van NLP-technieken voor het extraheren van relevante informatie zoals bedrijfsnamen, aandelen en aantallen. Vervolgens worden ML-modellen getraind om patronen en structuren te herkennen, en om de gegevens te classificeren en te structureren.
3. **Integratie:** Integreren van deze gegevens in een relationele databank die ontworpen is voor efficiënte opslag en toegang.

In de derde fase zal het systeem worden geëvalueerd op basis van enkele criteria: accuraatheid en efficiëntie en kwaliteit van de geëxtraheerde gegevens.

De resultaten van de gegevensextractie worden vergeleken met handmatig gecodeerde gegevens en de verwerkingstijd om de efficiëntie en nauwkeurigheid te evalueren.

De kwaliteit van de gegevens wordt gemeten door fouten en inconsistenties in de geëxtraheerde en genormaliseerde gegevens te vinden, naast de consistentie en volledigheid van de gestandaardiseerde gegevens.

Na evaluatie van het proof-of-concept systeem, worden de bevindingen gepresenteerd en aanbevelingen gedaan voor verdere verbeteringen en mogelijke toepassingen. Dit kan ook aanbevelingen omvatten voor bredere implementatie, zoals integratie met andere financiële analysetools en verdere verfijning van de AI-modellen op basis van feedback en aanvullende gegevens.

Deze gestructureerde aanpak zorgt ervoor dat het proof-of-concept systeem effectief en efficiënt de historische 13F-meldingen kan verwerken,

waardoor de toegankelijkheid en analyse van historische financiële gegevens wordt verbeterd.

4. Verwachte resultaten, conclusie

Het verwachte resultaat van het onderzoek is een werkende proof-of-concept applicatie te ontwikkelen die AI-technologieën gebruikt, waaronder NLP en ML- technologieën, om alle 13f-meldingen van voor 2013 te standaardiseren en integreren in een relationele databank. De applicatie die wordt ontwikkeld moet de gegevens binnen ene acceptabele tijd extraheren en verwerken naar een uniform formaat en vervolgens naar een databank weg te schrijven. Het gevolg hiervan is dat de toegankelijkheid en analyse van de historische financiële gegevens worden verbeterd en vergemakkelijkt. Hierdoor kunnen onderzoekers met minder inspanning en kosten diepere inzichten verkrijgen in historische beleggingstrends en gemakkelijker voorspellende modellen maken.

Kortom, AI-technologieën zoals NLP en ML kunnen een machtige oplossing bieden bij het standaardiseren en normaliseren van historische 13F-meldingen. Het systeem zal automatisch de inconsistenties in dergelijke documenten op. Het daaropvolgende bewijs-of-concept systeem zal een waardevolle input zijn voor financieel onderzoek en investeringsanalyse en zal fungeren als basis voor toekomstige toepassingen in de analyse van historische financiële data-analyse en de ontwikkeling van voorspellende modellen.

Referenties

AWS. (2024). *What is the difference between structured and unstructured data?* Verkregen augustus 22, 2024, van https://aws.amazon.com/compare/the-difference-between-structured-data-and-unstructured-data/

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17), 42–45. https://doi.org/10.5120/14937-3507

IBM. (2024). *What is text-mining*. Verkregen augustus 22, 2024, van https://www.ibm.com/topics/text-mining

Kinter, P. (2024, februari 12). *Text mining: applications and techniques*. Verkregen augustus 22, 2024, van https://www.alexanderthamm.com/en/blog/text-mining-basics-methods-and-application-cases/

Securities, U., & Commission, E. (2023). *Frequently Asked Questions About Form 13F*. Verkregen augustus 22, 2024, van https://www.sec.gov/divisions/investment/13ffaq