TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Andmeteaduse õppekava

Brandon Loorits

Ettevõtete jätkusuutlikkuse eelanalüüs aastaaruannete põhjal Balti börsi ettevõtete näitel

Magistritöö (15 EAP)

Juhendajad: Mark Fišel, PhD
Lehar Oha

Tartu 2024

Ettevõtete jätkusuutlikkuse eelanalüüs aastaaruannete põhjal Balti börsi ettevõtete näitel

Lühikokkuvõte:
Tänapäeval on jätkusuutlikkus tuntud ka kui ESG (ingl *environmental*, *social and corporate governance*) ehk keskkondlik, sotsiaalne ja juhtimisalane vastutus. See on muutunud ettevõtete strateegilises juhtimises ja investeerimisotsustes keskseks mõisteks üle kogu maailma, sealhulgas Balti riikides. Finantsasutused ja investorid pööravad üha enam tähelepanu ettevõtete jätkusuutlikkusele [ZB21]. Antud magistritöö käsitleb jätkusuutlikkuse eelanalüüsi mudeli arendamist, kasutades selleks suuri keelemudeleid Balti börsil noteeritud ettevõtete aastaaruannete analüüsimiseks. Töö eesmärk on hinnata ettevõtete panust ESG valdkondades ja pakkuda analüütikutele tööriistu suurte andmemasside tõhusamaks töötlemiseks. Uuringus keskendutakse suurtele keelemudelitele, nagu GPT-4 Turbo, mis võimaldavad automaatselt genereerida ESG eelanalüüsi küsimustikule vastuseid, vähendades nii manuaalse töö hulka kui ka tõstes analüüsi efektiivsust. Loodud prototüüp võimaldab jätkusuutlikkuse spetsialistidel kiiremini ja täpsemini hinnata ettevõtete jätkusuutlikkust ilma aastaaruandeid detailideni läbi lugemata.
Võtmesõnad: LLM, Keeleteadus

CERCS: P176 Tehisintellekt

Joonis 1. Graafiline lühikokkuvõtte eesti keeles.

Preliminary Analysis of Corporate Sustainability Based on Annual Reports: A Case Study of Companies on the Baltic market

Abstract:
Sustainability, also known as ESG (Environmental, Social, and Corporate Governance), has become a central concept in corporate strategic management and investment decisions worldwide, including in the Baltics. Financial institutions and investors are increasingly focusing on the sustainability of the companies they invest in [ZB21]. This master's thesis explores the development of a preliminary sustainability analysis tool using large language models to analyze the annual reports of companies listed on the Baltic stock exchange. The aim is to assess companies' contributions to ESG fields and provide analysts with tools for more efficiently processing large volumes of data. The study focuses on large language models, such as GPT-4 Turbo, which enable the automatic generation of answers to ESG pre-analysis questionnaires, thereby reducing manual labor and increasing analysis efficiency. The prototype developed allows sustainability specialists to more quickly and accurately assess corporate sustainability without the need to meticulously read through annual reports.

Keywords: LLM, Linguistics

CERCS: P176 Artificial Intelligence

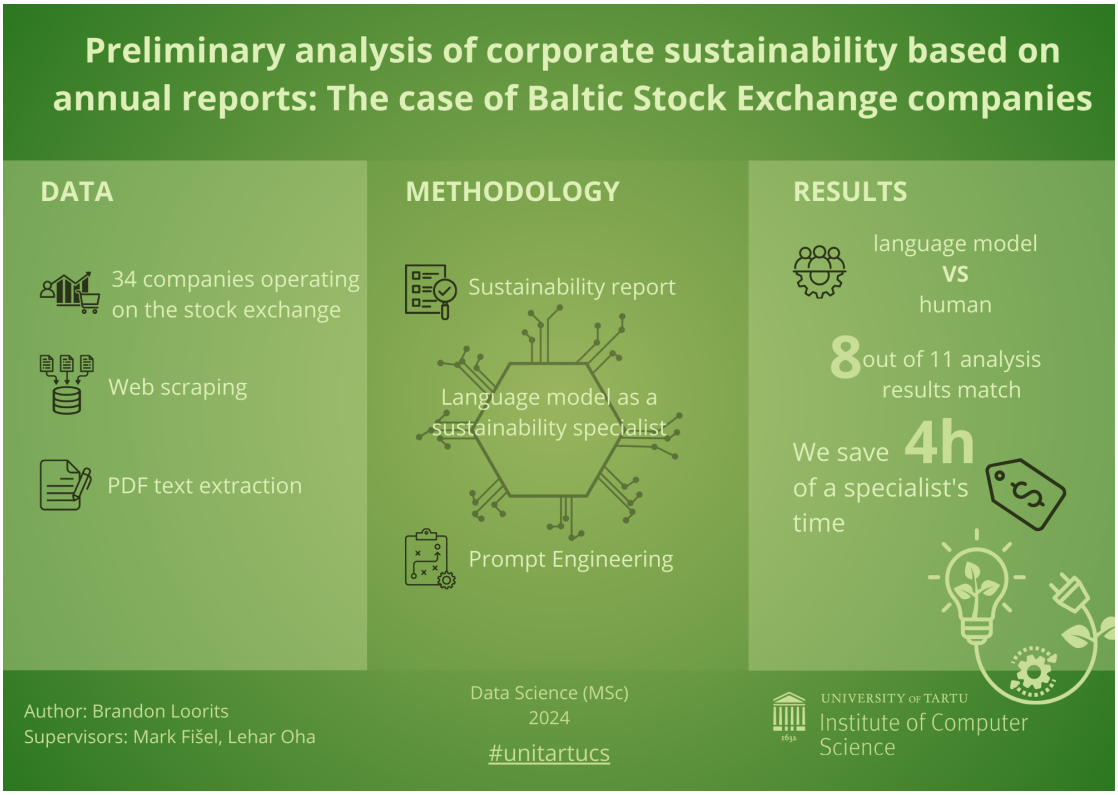


Figure 2. Graphical abstract in english.

Sisukord

1 Sissejuhatus

Jätkusuutlikkus on tänapäeval keskne mõiste ettevõtete strateegilises juhtimises ja investeerimisotsustes üle maailma, sealhulgas Balti riikides, kus finantsasutused ja investorid on üha enam hakanud keskenduma sellele, kui jätkusuutlikud on ettevõtted, kuhu nad oma kapitali suunavad. Uuringud näitavad, et ligikaudu 81% Balti riikide finantsturgude osalistest kasutavad keskkonnaalaseid, sotsiaalseid ning juhtimisalaseid andmeid investeeringute hindamisel, mis on märkimisväärne osakaal ja rõhutab jätkusuutlikkuse kasvavat rolli investeerimisprotsessis [ZB21].

Sellises kontekstis on oluline arendada vahendeid ja meetodeid, mis aitavad ettevõtetel ja finantsanalüütikutel hinnata jätkusuutlikkuse aspekte tõhusamalt. Selliste vahendite ja meetodite arendamine võimaldab automatiseerida jätkusuutlikkuse analüüsi etappe kiirendades seejuures üldist tööprotsessi.

Käesoleva magistritöö eesmärk on luua prototüüp jätkusuutlikkuse eelanalüüsi tegemiseks, kasutades suurt keelemudelit GPT-4 Turbo, mis aitab analüüsida ja töödelda suurtes kogustes aastaaruandeid. See on oluline, kuna traditsiooniline lähenemine nõuab, et jätkusuutlikkuse spetsialistid loeksid läbi arvukalt aastaaruandeid, mis on aeganõudev ja ressursimahukas ülesanne. Prototüübi väljatöötamise tulemusel lüheneb ka kogu jätkusuutlikkuse hinnangu koostamise protsess. Protsessi lühendamine võimaldab jätkusuutlikkuse spetsialistil keskenduda üldanalüüsi koostamisele.

Magistritöö autor kasutas prototüübis suurt keelemudelit, kuna on täheldatud, et suurte keelemudelite kasutamine, sealhulgas GPT-4, on näidanud olulisi edusamme struktureeritud andmete ekstraheerimisel ettevõtete ESG aruannetest. Artiklis "ESGReveal: An LLM-based approach for extracting structured data from ESG reports"kirjeldatakse, kuidas GPT-4 suudab saavutada 76,9% täpsust andmete ekstraheerimisel ja 83,7% täpsust andmete analüüsimisel [ZSC+23]. Need tulemused kinnitavad LLM-ide potentsiaali ESG andmete analüüsimisel ja parandamisel, mis on ülioluline tööriist investoritele ja korporatiivsetele finantsasutustele.

Magistritöö käigus koguti andmed Nasdaq-i veebilehelt, kust võeti kõigi Balti põhinimekirja kuuluvate ettevõtete aastaaruanded. Need aruanded töödeldi tekstikujule, misjärel vastav tekstikujuline sisend anti suure keelemudeli GPT-4 Turbo käsutusse. Selle mudeli abil genereeriti vastused jätkusuutlikkuse küsimustikule, mida tavaliselt täidavad jätkusuutlikkuse spetsialistid. Spetsialistide tulemusi võrreldi seejärel keelemudeli poolt antud vastustega, et hinnata mudeli tõhusust ja täpsust jätkusuutlikkuse aspektide hindamisel.

Peamised uurimisküsimused, millele käesolev magistritöö vastuseid otsib, on järgmised:

1. Kui tõhusalt suudab suur keelemudel GPT-4 Turbo analüüsida ja tõlgendada jätkusuutlikkuse andmeid võrreldes traditsiooniliste meetoditega?

2. Millised on suurte keelemudelite kasutamise võimalikud edasiarendused ja parandused jätkusuutlikkuse andmete analüüsimisel?

Magistritöö koosneb viiest peatükist. Esimeses peatükis selgitatakse jätkusuutlikkuse põhimõtteid, sealhulgas keskkondliku, sotsiaalse ja juhtimisalase vastutuse aspekte ning samuti kirjeldatakse jätkusuutlikkuse raamatupidamisstandardite nõukogu metodoloogiat. Teises peatükis kirjeldatakse, kuidas andmeid koguti Nasdaq-i lehelt, puhastati ja töödeldi tekstikujule. Kolmandas peatükis tutvustatakse uurimuse eesmärki ja olulisust ning selgitatakse suuri keelemudeleid, sealhulgas GPT-4 Turbo mudelit ning lähtelause sõnastamise protsessi. Neljandas peatükis käsitletakse saadud tulemusi ja võrreldakse neid oodatud vastustega. Viies peatükk keskendub tuleviku töö võimalustele, sealhulgasESG aruannete regulatsioonide ja andmemahtude suurendamisele ning päringuga täiendatud generatsiooni mudelite kasutamisele. Tähtis on viidata, et käesolevas magistritöös on kasutatud tehisintellekti keeleliselt teksti korrastamiseks ja vormistamiseks [Ope22].

2 Jätkusuutlikkuse põhimõtted

Käesoleva magistritöö eesmärgiks on luua prototüüp tööriistast, mis aitaks analüüsida ettevõttete jätkusuutlikkust. Lahenduse tulemuseks peaks olema tööriista prototüüp, mis vastab jätkusuutlikkuse eelanalüüsi küsimustikule. Kuna jätkusuutlikus on väga lai mõiste, siis selgitab magistritöö autor järgnevates alapeatükkides jätkusuutlikkuse tähendust ja põhimõtteid.

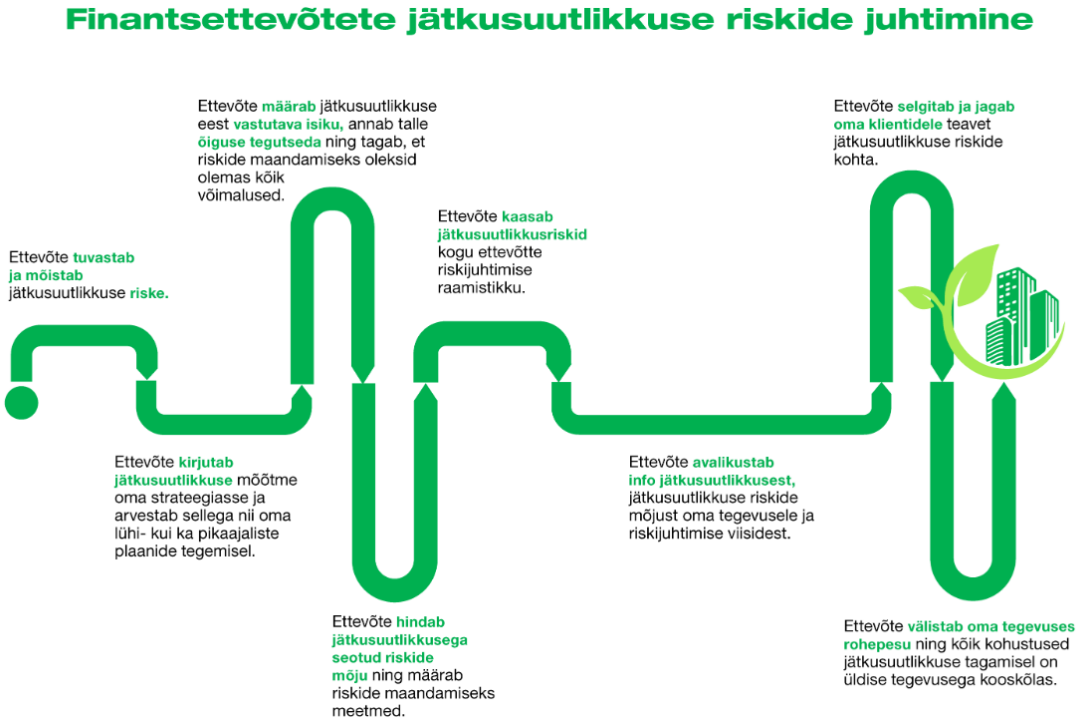
2.1 Keskkondlik, sotsiaalne ja juhtimisalane vastutus

Tänapäeval on peamiseks väljendiks jätkusuutlikkuse kohta ESG (ingl *environmental*, *social* and *coprorate governance*) ehk keskkondlik, sotsiaalne ja juhtimisalane vastutus. See kirjeldab ettevõtete tegevusi, mis võtavad arvesse keskkonnaalaseid, sotsiaalseid ja juhtimisega seotud aspekte. Need mõõdikud on olulised, kuna need aitavad investoritel ja teistel huvigruppidel hinnata ettevõtete jätkusuutlikkuse taset ja eetilisust [Arm20].

ESG tähtsus kasvab järjest enam, kuna tarbijad ja investorid nõuavad ettevõtetelt suuremat läbipaistvust ja vastutustundlikkust. Organisatsioonid, mis rakendavad ESG põhimõtteid, näitavad sageli paremaid finantstulemusi ja on investorite seas usaldusväärsemad. Lisaks on paljudes riikides, sealhulgas Euroopas, karmistunud ESG-alased regulatsioonid. Näiteks on S&P 500 ettevõtted, fond kuhu kuuluvad maailma 500 suurimat ettevõtet, alustanud ESG aruannete avalikustamist, mis oluliselt mõjutavad nende turuväärtust ja investorite otsuseid [AH20].

Jätkusuutlikkus on tänapäeval keskne mõiste ettevõtete strateegilises juhtimises ja investeerimisotsustes üle maailma, sealhulgas Balti riikides, kus finantsasutused ja investorid on üha enam hakanud keskenduma sellele, kui jätkusuutlikud on ettevõtted, kuhu nad oma kapitali suunavad. Uuringud näitavad, et ligikaudu 81% Balti riikide finantsturgude osalistest kasutavad keskkonnaalaseid, sotsiaalseid ning juhtimisalaseid andmeid investeeringute hindamisel, mis on märkimisväärne osakaal ja rõhutab jätkusuutlikkuse kasvavat rolli investeerimisprotsessis [ZB21].

Euroopa Liidus võeti vastu 14.detsembril 2022 kestlikkusaruandluse direktiiv, mis kohustab järjest enam ettevõtteid kestlikkuseandmeid koguma, analüüsima ja avaldama. [EU2]. Prantsusmaal on võetud kasutusele rangemad ESG regulatsioonid, mille eesmärk on kujundada vastutustundlikum mudel. Need regulatsioonid hõlmavad mitte ainult finantsaruandlust, vaid ka keskkonna- ja sotsiaalvaldkonna nõudeid, mille eesmärk on tagada ettevõtete ja nende partnerite vastavus jätkusuutlikkuse põhimõtetele [PdG22]. Need on näited sellest, et jätkusuutlikkus ei ole oluline ainult investeeringute tegemisel, vaid seda hindavad ka Euroopa suurimad institutsioonid. Näiteks on Eesti Finantsinspektsioon loonud teekardi jätkusuutliku investeerimise teekonnal olevatele ettevõtetele, mis on nähtav joonisel 3.



Joonis 3. Finantsinspektsiooni teekaart jätkusuutliku investeerimise teekonnal olevatele ettevõtetele[Ees24].

2.1.1 Keskkond (E)

Keskkondlik vastutus hõlmab ettevõtte mõju füüsilisele keskkonnale. See sisaldab tegevusi, mis mõjutavad kliimamuutusi, ressursside kasutust, jäätmete käitlemist ja looduslike elupaikade säilitamist. Ettevõtted, kes pühenduvad keskkonnaalasele vastutusele, rakendavad praktikaid, mis vähendavad nende ökoloogilist jalajälge ja edendavad jätkusuutlikkust [AL19].

2.1.2 Sotsiaalne vastutus (S)

Sotsiaalne vastutus käsitleb ettevõtte mõju ühiskonnale. See hõlmab töötajate heaolu, töötingimusi, kogukonna kaasamist ja ettevõtte mõju kohalikele kogukondadele. Sotsiaalselt vastutustundlikud ettevõtted püüavad parandada oma töötajate, tarbijate ja kogukondade elukvaliteeti, milles nad tegutsevad [PHS+22].

2.1.3 Juhtimine (G)

Juhtimine viitab ettevõtte juhtimisstruktuuridele ja -poliitikatele. See hõlmab läbipaistvust, juhtkonna vastutust, korruptsioonivastaseid meetmeid ja ettevõtte üldist juhtimist. Tugev juhtimine tähendab tõhusat juhtimist, mis austab aktsionäride õigusi ja tagab ettevõtte eetilise käitumise kõikides oma tegevustes [PHS+22].

2.2 Jätkusuutlikkuse raamatupidamisstandardite nõukogu

Järgnevas peatükis kirjeldatakse SASB (ingl *Sustainability Accounting Standards Board*) ehk jätkusuutlikkuse raamatupidamisstandardite nõukogu metodoloogiat kuna küsimused, millele püüab suur keelemudel vastata, põhinevad SASB metodoloogial.

SASB pakub raamistikku, mis põhineb väljatöötatud SASB standarditel, et organisatsioonid saaksid teha tööstuspõhiseid aruandeid jätkusuutlikkusega seotud riskide ja võimaluste kohta, mis võivad mõjutada ettevõtte rahavoogusid, finantseerimisvõimalusi või kapitalikulu lühikeses, keskmises või pikas perspektiivis. SASB standardid on välja töötatud läbi range ja läbipaistva protsessi, mis hõlmab tõenduspõhist uurimistööd, laialdast osalust ettevõtetelt, investoritelt ja ala ekspertidelt ning iseseisva SASB-i järelevalvet ja heakskiitu [SAS].

2.2.1 Protsess ja standardid

SASB standardite eesmärk on aidata ettevõtetel standardiseerida aruandlust jätkusuutlikkuse küsimustes, mis on olulised iga tööstusharu jaoks. Need standardid hõlmavad keskkonnaalaseid, sotsiaalseid ja juhtimisalaseid teemasid, mis on tõenäoliselt finantsiliselt olulised. SASB on välja töötanud juhendid 79 tööstusharu jaoks 11 sektoris, aidates sellega ettevõtetel keskenduda kõige asjakohasematele jätkusuutlikkuse küsimustele [FW20].

Samuti on ära määratud küsimused ja nende faktorid käesolevas magistritöös kasutatud andmetes. Nagu eelnevalt kirjeldatud, siis need faktorid ja küsimused põhinevad SASB metodoloogias väljatöötatud standardite põhjal, mis olenevad tööstusharust, milles ettevõte tegutseb. Vastavad faktorid ja küsimused on sisendiks suurele keelemudelile süsteemikäsus.

2.2.2 Tähtsus ja mõju

SASB metodoloogia tähtsus seisneb selle võimes pakkuda selget ja tööstuspõhist raamistikku, mis võimaldab ettevõtetel jätkusuutlikkuse andmeid tõhusamalt integreerida oma finantsaruandlusega. Näiteks BlackRocki tegevjuht Larry Fink on rõhutanud, et ettevõtted peaksid avalikustama oma jätkusuutlikkuse algatusi vastavalt SASB tööstuse spetsiifilistele juhistele, mis aitab parandada läbipaistvust ja hõlbustab investeerimisotsuste tegemist. Sellised algatused on muutumas üha olulisemaks, kuna üha rohkem

investoreid ja reguleerijaid nõuavad ettevõtetelt jätkusuutlikkuse aspektide arvestamist nende finantsaruannetes [FW20].

SASB lähenemine on aidanud kaasa sellele, et jätkusuutlikkus ei ole enam ainult avalike suhete tööriist, vaid strateegiline element, mis toetab ettevõtete pikaajalist väärtuse loomist ja riskijuhtimist. Selline strateegiline lähenemine jätkusuutlikkusele võimaldab ettevõtetel mitte ainult järgida seadusandlikke nõudeid, vaid ka parandada oma mainet ja finantstulemusi, muutes need investeerimisotsuste tegemisel atraktiivsemaks [SAS].

3 Andmed

Järgnevas peatükis kirjeldab magistritöö autor andmete kogumise protsessi, mille eesmärgiks oli Balti börsil olevate ettevõtete aastaaruannete kogumine, tekstistamine, puhastamine ja töötlemine. Selle tulemusel saadi suure keelemudeli tektsiline sisendfail.

3.1 Andmete kogumine

Andmete kogumine toimus *https://nasdaqbaltic.com/et* Balti börsi veebilehelt. Uurimiseks valiti kõik 34 Balti börsi põhinimekirja kuuluvat ettevõtet. Andmed koguti veebikraapimise teel, tehes päringuid Nasdaq-i veebilehele. Veebikraapimisprotsessis järgiti viisakusreegleid, et vältida lehe ülekoormamist. Päringute tegemisel piirati nende kiirust, et mitte serverit üle koormata, ja määrati päringu päises selge identiteet kasutades 'User-Agent' väärtust 'my_crawler (brandon.loorits@ut.ee) / for_study_purpose'. See lähenemine võimaldas selgelt näidata päringute eesmärki ja päritolu, vähendades riski, et päringuid blokeeritakse kui potentsiaalne rünnak.

Esmalt, et leida kõik Balti börsi põhinimekirjas olevad ettevõtted, tehti päring *https://nasdaqbaltic.com/statistics/et/shares* leheküljele, kust on leitavad Balti börsi põhinimekirjas olevad ettevõtted. Seejärel koguti kõigi nende ettevõtete üksikasjalikuma alamleheni viivad lingid BeautifulSoup [Ric23] teegi abil vastavaid märgendeid sõeludes ja otsides. Alamlehtedele tehti omakorda päringud ning leiti lingid, mis viivad alamleheni, kus on esitatud kõik ettevõtte aruanded. Aruannete alamlehelt otsiti märgendite abil välja viimati esitatud aastaaruanded ja kestlikkuse aruanded.

Andmete kogumise esmaseks allikaks olid ettevõtete aastaaruanded PDF formaadis. Kui PDF formaadis aastaaruandeid ei olnud saadaval, kasutati alternatiivina tihendatud faile, mis sisaldasid aastaaruandeid XHTML formaadis. Andmekogumise protsessis koguti kokku 33 ettevõtte aastaaruanded tekstikujul. Üks PDF formaadis aastaaruanne oli rikutud, kuna sellest ei olnud võimalik tekstilist sisu ekstraktida. 33 ettevõtte aastaruandest 26 puhul leiti PDF kujul fail, millest üks ei vastanud nõuetele. 8 ettevõtte aastaruande andmed loeti veebist maha XHMTL kujul. 2 ettevõtte puhul leiti ka ESG aruanne. PDF-failid ja XHTML-failid salvestatakse eraldi kaustadesse kuna need vajavad erinevaid tekstitöötluse protsesse.

Valideerimisandmed saadi ettevõte käest, kellega koostööd tehti. Samuti ka eelanalüüsi küsimused. Andmed koguti manuaalsel viisil ettevõtte poolt väljatöötatud jätkusuutlikkuse hindamiseks loodud rakendusest. Valideerimiseks saadi andmed 11 Balti börsi põhinimekirjas oleva ettevõtte kohta. Magistritöö autor märgib, et täpseid küsimusi ja valideerimisandmeid avalikustada ei saa kuna need on ainult ettevõtte siseseks kasutamiseks.

3.2 Andmete töötlemine ja tekstistamine

PDF ja XHTML failidele rakendati erinevaid tekstitöötlus protesse ning seetõttu olid salvestatud failid erinevatesse kaustadesse. Andmete lugemiseks erinevatest kaustadest kasutati teeke glob [Fou23a] ja os [Fou23b]. Teekide abil oli võimalik leida vajalik kaust, kust andmeid hakatakse maha lugema. Glob teegi abil sai leida nimekirja failidest, mis selles kaustas leiduvad. Nimekirja alusel oli võimalik leida faili nimi ja selle alusel otsustada, millist tekstitöötlusprotsessi rakendada.

Teksti ekstraheerimiseks kogutud PDF-failidest kasutati PyMuPDF [AS23] teeki, mis on tuntud oma võime poolest töödelda visuaalselt rikaste dokumentide tekstikihte. Teek võimaldab eraldada PDF-failidest nii tabeleid, pilte kui ka teksti. XHTML-failide andmete kättesaamiseks rakendati BeautifulSoup teeki, mis on efektiivne tööriist HTML ja XML dokumentide sõelumiseks ja andmete ekstraheerimiseks. Täiustamaks ekstraheerimise protsessi, kasutati regulaaravaldiste (regex) mustreid, et identifitseerida ja välja lõigata spetsiifilised andmeplokid. Andmeplokkide pikkus leiti sisukorrast vastavate pealkirjade otsimisel, kust leiti soovitud lehekülje number kuni milleni info aastaruandest eraldati.

Aastaaruannetest olulise teabe leidmiseks ja struktureerimiseks kasutati PyMuPDF ja BeautifulSoup teekide funktsionaalsusi. Pärast relevantsete lehekülgede tuvastamist ekstraheeriti tekstid kuni konsolideeritud finantsandmeteni. Erinevate aastaaruannete varieeruv struktuur ja pealkirjad tõid kaasa vajaduse mõningatel juhtudel lehekülje numbrid käsitsi ette anda, et tagada täpne andmete kogum. XHTML formaadis dokumentide puhul rakendati spetsiifilisi märgendeid, mis võimaldasid vajaliku sisu tõhusalt välja lugeda.

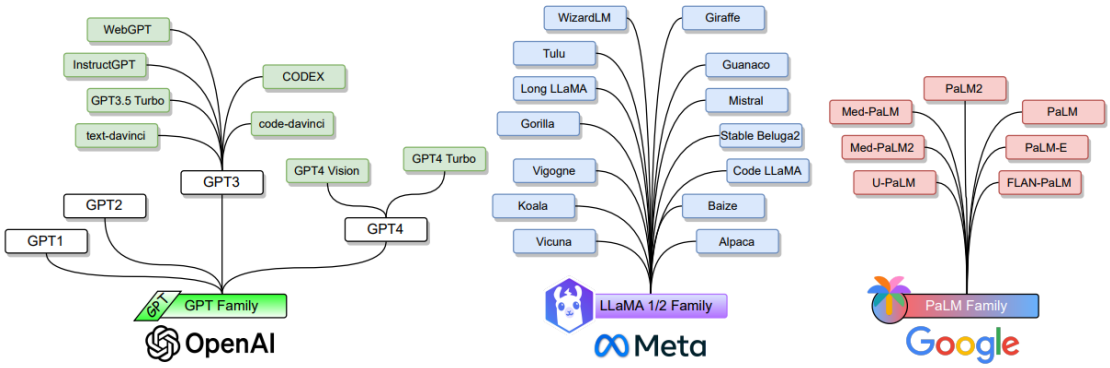
Uurimisprotsessi käigus ilmnenud väljakutsed, nagu rikutud PDF-failid ja erinevad dokumendiformaadid, nõudsid andmete eraldamise protsessi kohandamist. Selleks rakendati alternatiivseid formaate ja kohandati ekstraktsioonimeetodeid, et ületada mainitud takistused ja tagada kogutud andmete täielikkus ning usaldusväärsus. Täielikkuse saavutamiseks kasutati erinevaid tekstipuhastus meetodeid nagu tühemike eemaldamine, üleliigsete tühikute eemaldamine, tükeldatud sõnade liitmine jne. Samuti rakendati paindlikke andmetöötlusmeetodeid, et kohandada andmeeraldusprotsesse vastavalt dokumentide eripäradele, tagades sellega, et kõik relevantne informatsioon saadi kätte ja töödeldi nõuetekohaselt. Need meetmed olid vajalikud, et säilitada magistritöö kvaliteet ja täpsus, võimaldades seeläbi saavutada uurimiseesmärkide täitmist.

4 Suured keelemudelid

Suured keelemudelid, eriti GPT-seeria, on saavutanud märkimisväärse tähtsuse loomuliku keele töötlemise valdkonnas. Järgmistes peatükkides keskendub magistritöö autor suurte keelemudelite olulisusele, GPT-seeria mudelitele ning lähtelause sõnastamisele. Peatüki eesmärk on selgitada, kuidas GPT-4 mudelit kasutati jätkusuutlikkuse analüüsis magistritöös, ning tutvustada lähtelause sõnastamise tehnikaid selle rakendamisel. Lisaks selgitatakse, kuidas lähtelauseid kohandati vastavalt analüüsi eesmärkidele ning millised olid olulised komponendid lähtelause struktuuris ja formaadis.

4.1 Suurte keelemudelite olulisus

Keelemudelid on NLP (ingl *Natural Language Processing*) ehk loomuliku keele töötlemise nurgakivi, kasutades matemaatilisi meetodeid keeleliste reeglite üldistamiseks, ennustamiseks ja genereerimiseks. Keelemudelite areng on olnud viimastel aastakümnetel märkimisväärne alates esialgsetest SLM (ingl *Statistical Language Model*) ehk statistilistest keelemudelitest kuni praeguste LLM-ideni (ingl *Large Language Model*) ehk suurte keelemudeliteni. Suurte keelemudelite puhul eristatakse kolme perekonda - PaLM, LLaMa ja GPT, mille mudelid on loetletud joonisel 4. Viimane neist on saanud kõige rohkem tähelepanu viimastel aastatel kuna on näidanud parimaid tulemusi. GPT kasutusala on laienenud mitmetesse erinevates valdkondadesse nagu meditsiin, ärianalüütika, juriidika jne [CNW+24].



Joonis 4. Suurte keelemudelite perekonnad[MMN+24].

Suured keelemudelid, mida illustreerivad näiteks GPT-seeria mudelid, töötavad põhimõtetel, et mõista ja luua inimese sarnast keelt. Esiteks võtavad nad sisendiks sümbolite (sõnade või tokenite) jadu ning kodeerivad need numbrilisteks vektoriteks. Need vektorid esindavad sõnade tähendusi kokkusurutud ruumis, saavutades seda näiteks ühe-soojuse

(ingl *one-hot*) kodeerimise ja sissekukkumis (ingl *embedding*) maatriksite abil. Oluliselt arvestavad LLM-id ka sõnade järjestust ja positsioone lauses, kasutades positsiooni põhist kodeerimist, tagamaks konteksti täpne mõistmine. Enesetähelepanu (ingl *self-attention*) mehhanismid võimaldavad mudelil keskenduda olulistele sisendi osadele, mida veelgi täiendatakse mitmik tähelepanuga (ingl *multi-head attention*) paralleelse töötlemise jaoks. Ennustatavatele sõnadele juurdepääsu vältimiseks ennustamise ajal kasutatakse maskeerimistehnikat. Lisaks suurendavad pärilevivõrgud ja normaliseerimistehnikad mudeli väljendusjõudu ja stabiilsust. Pärast mitmeid töötlusetappe väljastab mudel tõenäosused igale sõnale sõnavaras, võimaldades valida kõige tõenäolisema väljundi. Need põhimõtted üheskoos võimaldavad LLM-idel tõhusalt mõista ja luua loomulikku keelt [CNW+24].

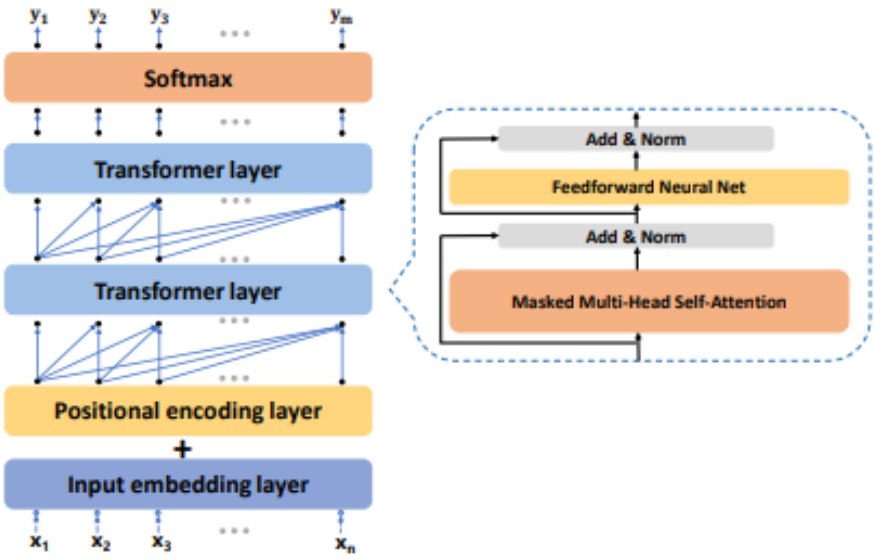
GPT seeria, mis on treenitud ulatuslike tekstiandmetega ja demonstreerivad võimet genereerida inimtasemel tekste ja täita keelepõhiseid ülesandeid erakordse täpsusega. Need arengud on oluliselt parandanud erinevaid töö- ja eluaspekte, kuigi LLMide laiem mõistmine võib praktikutel olla piiratud, eriti neil, kellel puudub taustteadmised NLP-s [CNW+24].

4.2 GPT seeria

Keelemudelite arengus on viimastel aastakümnetel toimunud olulisi muutusi, eriti pärast transformermudelite ilmumist, mis on viinud suurte keelemudelite arenguni. GPT (ingl *Generative Pre-trained Transformer*) ehk generatiivsete eeltreenitud transformerite mudelite perekond, mille on välja töötanud OpenAI, on üks märkimisväärne näide sellistest mudelitest. GPT mudelid, sealhulgas GPT-3 ja hiljuti välja antud GPT-4, on tuntud oma suurte parameetrite arvu ja sügava õppevõime poolest, mis võimaldab neil sooritada keerukaid keelelisi ülesandeid ja genereerida inimtasemel teksti[MMN+24]. Transformertüüpi GPT mudeli konseptuaalne arhitektuur ja selle kihid on nähtavad joonisel 5.

GPT-3 oli esimene, mis oluliselt ületas eelmisi mudeleid. Selle võimekus ilmneb GPT-3 erakordses mitmekülgsuses, suutes genereerida teksti, mis võib järgida väga erinevaid juhiseid ja toime tulla mitmesuguste ülesannetega ilma täiendava kohandamiseta. GPT-4, mis on veelgi võimsam, toob kaasa suuremaid võimalusi ja on treenitud veelgi suurema andmekogumi peal, võimaldades keerukamaid keelelisi toiminguid ja paremat mõistmist. Need mudelid demonstreerivad, kuidas LLM-id võivad ületada varasemate mudelite piirangud, pakkudes paremat keelelist mõistmist ja generatiivset võimekust, mis on olulised atribuudid paljudes rakendustes, alates automatiseeritud klienditeenindusest kuni andmete analüüsini [MMN+24].

GPT-4 Turbo, edasiarendatud versioon GPT-4 mudelist, on välja töötatud eriti suurte parameetrite arvu ja tõhusama vastuste genereerimise võimekusega. Kuigi GPT-4 Turbo ei ole oma olemuselt deterministlik, mis tähendab, et selle vastused võivad sama sisendi korral erineda, on võimalik saavutada deterministlikumaid tulemusi, ka-



Joonis 5. GPT mudeli konseptuaalne arhitektuur[CNW+24].

sutades näiteks järjepidevaid suunavaid strateegiaid. Viimased uuringud on näidanud GPT-4 Turbo mitmekülgset rakendamist tarkvaraarenduse ülesannete automatiseerimisel, näiteks nõuete loomise puhul, mis näitab selle olulist mõistmist eesmärgipõhistes modelleerimistes[SSM+23].

Käesolevas magistritöös kasutatakse suurte keelemudelite GPT seeriast GPT-4 Turbo mudelit. Keelemudelit kasutatakse eelanalüüsi küsimustele vastamiseks, mis on sisendiks üldisele jätskusuutlikuse analüüsi raportile. GPT-4 Turbo mudelile antakse aastaaruanne sisendinfona ning samuti antakse kasutaja poolt automaatselt küsimused, millele suur keelemudel sisendinfo ja treenitud andmete põhjal vastata püüab. Mudelile ligipääs tagati Tartu Ülikooli poolt, mille õigused anti juhendaja poolt.

Selleks, et sooritada järgnevates peatükkides kirjeldatud päringuid kasutati OpenAI teeki. Teegis leiduvad kõik vajalikud funktsioonid - päringute defineerimiseks, sooritamiseks ning vastuste intepreteerimiseks. Päringus määrati ära süsteemikäsk, milleks oli lähtelause ning kasutajakäsk, milleks oli aastaaruanne puhastatud kujul ilma konsolideeritud osadeta. Tuleb märkida, et kuna OpenAI rakendusliidese kaudu päringute tegemine on tasustatud sisendsümbolite põhjal, siis sellel eesmärgil piirati aastaaruande pikkust ja sisu.

4.3 Lähtelause sõnastamine

Eelmises peatükis kirjeldas magistritöö autor kui võimekas GPT-4 mudel olla võib. Kasutades sellel järjepidevaid suunavaid strateegiaid, siis on võimalik saavutada veel-

gi paremad tulemused. Üheks selliseks suunavaks strateegiaks võibki olla lähtelause sõnastamine (ingl *Prompt Engineering*). Sellist suunavat strateegiat kasutati ka selle magistritöö puhul. Järgnevas peatükis selgitabki magistritöö autor, kuidas aitab tulemuste täiustamisele kaasa lähtelause sõnastamine, millised on lähtelause sõnastamise võimalused ning milliseid strateegiad neist kasutati.

Lähtelause sõnastamine on protsess, kus kujundatakse spetsiifilisi sisendeid keelemudelile, et suunata ja optimeerida selle vastuseid kindlas kontekstis. See lähenemine on muutunud oluliseks, kuna võimaldab keelemudelitel, nagu GPT-4, luua asjakohasemaid ja täpsemaid vastuseid vastavalt kasutaja vajadustele. Prompt engineering mängib kriitilist rolli suurte keelemudelite rakendamisel erinevates valdkondades, aidates parandada nii täpsust kui ka kasulikkust teabe genereerimisel [JW].

Selles kontekstis on lähtelause sõnastamine kasulik, kuna see võimaldab kasutajatel juhtida mudeli suhtlemist ja suurendada selle efektiivsust spetsiifilistes rakendustes. Tehnika annab kasutajatele tööriista, et kohandada ja täpsustada, kuidas keelemudelid teavet töötlevad ja esitavad, muutes need tõhusamaks abilisteks otsuste tegemisel ja analüütilistes ülesannetes [JW].

Magistritöös kasutati lähtelausestamise tehnikat, et kujundada kontekstipõhiseid lähtelauseid. See strateegia aitas kaasa magistritöö eesmärgile – parandada masinõppe mudelite võimekust mõista ja töödelda jätkusuutlikkusega seotud andmeid. Täpsemalt, kontekstipõhised lähtelaused aitasid suunata mudeli tähelepanu asjakohastele aspektidele ja tagada, et genereeritud vastused oleksid tihedamalt seotud spetsiifiliste jätkusuutlikkusega seotud küsimustega. Samuti määratleti ära lähtelauses vastuse struktuur ja tüüp.

Esmalt testiti lähtelauset, mis on nähtav joonisel 6. Joonisel 6 nähtuva lähtelause puhul defineeriti mudeli jaoks ära metoodika, mille alusel peab küsimustele vastama. Samuti määrati lähtelauses kontekst, millele fookus tuleb sunnata, kuna sisendtekst, mis mudelile anti sisaldas ka osasid, mis ei olnud otseselt jätkusuutlikkuse aspektidega seotud. Tähelepanu suunamiseks anti ette teema, millele keskenduda tuleb. Magistritöö autor valis konteksti piiramiseks just teema, mitte faktorid ega lihtsalt jätkusuutlikkusega seonduva piirangu. Valiku põhjuseks oli, et kontekst oleks piisavalt määratletud ning mudel suunaks tähelepanu just kõige tähtsamatele osadele, kuid samas ei oleks kontekst liiga üldine.

Lähtelause oluliseks osaks oli ka ära defineerida vastuse struktuur ja formaat, et saada deterministlikud ja masinloetavad vastused, mitte kirjeldavad vastused nagu generatiivsed mudelid originaalselt annavad. Lähtelause osaks oli ka ülesande kirjeldus, mida mudel tegema peaks ehk vastama küsimustikule ning igale küsimusele valima ainult ühe valikvastuse variantidest.

Lähtelause esimese versiooni puhul oli suureks puuduseks, et paljudele küsimustele ning eraldi seisvatele faktoritele vastuseid ei saadud. Seetõttu täiendati lähtelauset vajalike osadega, milleks oli täpsem lähtelause kirjeldus, et mudelil ei oleks võimalik jätta

Lähtelause 1:

1. Act as a sustainability specialist, who is filling in the pre-analysis questions based on the SASB methodology.
2. Focus on finding the relevant parts of text based on the topic and then try to answer.
3. Choose one of the answer options, which is most suitable according to the provided report and give only the answer order number.
4. Give only one answer for every factor in format 'factor:answer number', example: 'Employee Safety:1', nothing more - only factor and answer.

Joonis 6. Lähtelause esimene versioon

vastamata ühtegi küsimust ning igale küsimusele peab vastus kindlasti antud olema. Ülejäänud lähtelause osad jäeti samaks. Lähtelause teine versioon on nähtav joonisel 7.

Lähtelause 2:

1. Act as a sustainability specialist, who is filling in the pre-analysis questions based on the SASB methodology.
2. Focus on finding the relevant parts of text based on the topic and then try to answer.
3. Choose one of the answer options, which is most suitable according to the provided report and give only the answer order number.
4. Ensure that for each question and factor, one precise answer must be provided. It is not possible to leave any question or factor unanswered.
5. Give only one answer for every factor in format 'factor:answer number', example: 'Employee Safety:1', nothing more - only factor and answer.

Joonis 7. Lähtelause teine versioon

Eelnevalt kirjeldatud lähtelaused olid aluseks päringu koostamisel. Kuid päringule lisati ka küsimused, vajadusel ka faktorid kui küsimuse puhul oli vajalik faktoreid hinnata eraldiseisvalt ning samuti valikvastused. Vastavad küsimused, faktorid ning valikvastused loeti sisse eelnevalt loodud andmetabelist, mis on loodud magistritöö autori poolt valideerimiseks saadud andmete põhjal. Vastavate lähtelausete, küsimuste, faktorite ja valikvastuste alusel loodi lõplikud lähtelaused.

Lähtelaused erinesid kahel juhul - kui vajalik oli vastata ainult küsimusele või küsimusele seejuures vastavalt igale faktorile erladiseisvalt. Joonisel 8 on nähtav lähtelause

versioon küsimuse puhul ning joonisel 9 nähtav lähtelause faktoritega küsimuste puhul. Faktorid olid määratletud valideerimisandmete poolt, mis sõltuvad sellest, millises sektoris ettevõte tegutseb.

Lähtelause küsimuste puhul:
Act as a sustainability specialist, who is filling in the pre-analysis questions based on the SASB methodology.
Focus on finding the relevant parts of text based on the topic and then try to answer.
Choose one of the answer options, which is most suitable according to the provided report.
Ensure that for each question and factor, one precise answer must be provided. It is not possible to leave any question or factor unanswered.
Give only one answer in format 'question:answer number', example: 'What number is best?:1', nothing more - only question and answer.
Topic: Renewable energy usage
Questions:

- How does the company measure its renewable energy usage and set goals for improvement?

Answers:

1. The company does not track or report any data related to renewable energy usage.
2. Renewable energy usage data is collected informally but not used to set specific improvement goals.
3. The company sets occasional renewable energy usage goals, but they are not regularly reviewed or updated.
4. Renewable energy usage is measured and improvement goals are set annually, but they are not publicly disclosed.
5. The company tracks renewable energy usage comprehensively, sets ambitious improvement goals, and publicly discloses progress.
6. Renewable energy usage metrics and goals are independently verified by a reputable third party.

Joonis 8. Lähtelause ainult küsimuse puhul

Lähtelause faktorite puhul:
Act as a sustainability specialist, who is filling in the pre-analysis questions based on the SASB methodology.
Focus on finding the relevant parts of text based on the topic and then try to answer.
Choose one of the answer options, which is most suitable according to the provided report and give only the answer order number.
Give only one answer for every factor in format 'factor:answer number', example: 'Employee Safety:1', nothing more - only factor and answer.
Question:

- How does the company prioritize and support employee well-being and work-life balance within its organizational culture?

Factors:

- Health and Safety Policies
- Flexible Work Arrangements
- Mental Health Support Programs
- Employee Benefits and Perks
- Professional Development Opportunities
- Diversity and Inclusion Initiatives

Answers:

1. The company does not demonstrate any clear commitment to prioritizing employee well-being or work-life balance.
2. While flexible work arrangements exist, they are not structured or formalized within the company's policies.
3. The company provides mental health support programs, but they are not widely communicated or utilized by employees.
4. Employee benefits and perks are available, but there is limited emphasis on promoting work-life balance through these offerings.
5. Professional development opportunities are provided, with some consideration for balancing work responsibilities and personal growth.
6. Diversity and inclusion initiatives are prioritized, contributing positively to employee well-being, but there is room for improvement in work-life balance support.

Joonis 9. Lähtelause faktorite puhul

5 Tulemused

Tulemuste peatükis esitatakse magistritöö käigus saavutatud uurimistulemused, mis hõlmavad suure keelemudeli GPT-4 Turbo poolt genereeritud vastuste võrdlemist jätkusuutlikkuse spetsialistide poolt antud vastustega. Analüüsitakse, kuidas keelemudel suutis interpreteerida ja vastata keerukatele küsimustele, mis põhinevad finantsasutuse poolt esitatud aastaaruannetel. See peatükk selgitab, millises ulatuses keelemudeli vastused kattuvad professionaalsete analüütikute hinnangutega ning tuuakse välja nii ühtivused kui erinevused. Lisaks käsitletakse, milliseid praktilisi järeldusi on võimalik saadud tulemustest teha. Selgitatakse ka, kuidas tulemused võivad mõjutada edasist tööd suurte keelemudelite rakendamisel finantsanalüüsis ja jätkusuutlikkuse hindamises.

Analüüsi koostamisel kasutati 581 andmepunkti, mis koosnesid erinevatest küsimustest ja faktoritest. Faktorid erinesid vastavalt valdkonnale, milles ettevõtte tegutseb. Iga küsimus anti GPT-4 Turbo mudelile eraldiseisvalt ja anti ette ka faktorid, mida mudel peab hindama. Hindamine tähendab, et mudel pidi valima valikvastuste seast kõige enam sobivama vastusevariandi. Kõik tulemused salvestati andmetabelisse ning salvestati ka exceli tabelisse hilisema analüüsi koostamiseks. Kui lähtelause esimene versioon andis vastused 539 juhul 581-st, siis teise versiooni puhul saadi vastused kõigile küsimustele. Seetõttu kirjeldab magistritöö autor lähtelause teise versiooniga saadud tulemusi.

Tabel 1 kajastab jätkusuutlikkuse spetsialisti ja GPT-4 Turbo mudeli poolt antud vastuste ühtivust lähtelause teise versiooni puhul, mis on nähtav jooniselt 7. Tabelist 1 nähtub, et lähtelause teise versiooni puhul ühtisid vastused 143 korral. Esimese lähtelause puhul leidus ka juhus, kus ükski vastus ei ühtinud, mis oli ettevõtte Auga group puhul.

Tabel 1. Vastuste ühtivus lähtelause teise versiooniga

Ettevõte	**Ei ühti**	**Ühtivad**	**Protsent (%)**
AUGA group	11	3	21.43
EfTEN Real Estate Fund	12	5	29.41
Enefit Green	57	15	20.83
Ignitis grupė	62	10	13.89
Merko Ehitus	51	7	12.07
Nordecon	50	8	13.79
Pieno žvaigždės	43	29	**40.28**
TKM Grupp	35	9	20.45
Tallinna Sadam	40	11	21.57
Tallinna Vesi	29	22	**43.14**
Vilkyškių pienin	48	24	**33.33**

Lähtelause teise versiooni kasutades leiti kõigi ettevõtete puhul ühtivaid vastuseid. Kõige rohkem ühtisid vastused ettevõtete Tallinna Vesi, Pieno žvaigždes ja Vilkyškiu

pienin puhul. Kõige vähem ühtisid aga etevõtete Ignitis grupe, Merko Ehitus ja Nordecon puhul. Kõige suurem protsent, mis ühtimiste puhul leiti, oli 43.14%, mis on peaaegu pooltel juhtudel.

Kuivõrd jätkusuutlikkuse hindamine on suhteline ja seda võib igaüks erinevalt hinnata, nagu näiteks ka kaks samas ettevõttes töötavat jätkusuutlikkuse spetsialisti, siis tuleks hindamisel vaadata ka tulemusi, mis erinevad ainult ühe võrra. See on vajalik selleks, et saada aru, kas tulemused on õiges suunas või on need radikaalselt valed. Need tulemused on nähtavad tabelist 2, kus on näidatud tulemused 1 punktilise veamääraga.

Tabelist 2 nähtub, et rohkem kui pooltel juhtudel on mudel hinnanud ettevõtete jätkusuutlikkust sarnaselt jätkusuutlikkuse spetsialistidele. Kõige kõrgemal juhul sarnanevad spetsialisti ja mudeli tulemused 86.27% ulatuses. Kuna 8 ettevõtte puhul 11-st on tulemused poolel juhul sarnased spetsialisti tulemustega, siis võib öelda, et mudelile veelgi täpsemate lähtelausete ja rohkemate andmete puhul oleks mudeli vastustel palju potensiaali. Meetodid, mis aitaks veelgi täpsemaid ja usaldusväärsemaid tulemusi saada kirjeldab magistritöö autor järgmises peatükis.

Tabel 2. Vastuste ühtivus lähtelause teise versiooniga - 1 puntkilise veamääraga

Ettevõte	**Ei ühti**	**Ühtivad**	**Protsent (%)**
AUGA group	10	4	28.57
EfTEN Real Estate Fund	6	11	64.71
Enefit Green	28	44	61.11
Ignitis grupė	61	11	15.28
Merko Ehitus	21	37	63.79
Nordecon	31	27	46.55
Pieno žvaigždės	18	54	**75.00**
TKM Grupp	26	18	40.91
Tallinna Sadam	22	29	56.86
Tallinna Vesi	7	44	**86.27**
Vilkyškių pienin	14	58	**80.56**

Tabel 3 kirjeldab tulemusi, mis näitavad kui mitmel juhul oli finantsasutuse jätkusuutlikkuse spetsialisti poolt antud hinnang kõrgem kui GPT-4 mudeli poolt. On oluline märkida, et vastuste võrdlemisel võeti andmestikust välja juhud kui vastused ühtisid. Tabelist 3 nähtub, et 7 korral hindas jätkusuutlikkuse spetsilist kõrgemalt ettevõtte jätkusuutlikkuse aspekte kui GPT-4 Turbo mudel. Selline olukord võib lihtsasti tekkida kuna spetsialistidel tekib tahtmatult soov enda portfellis olevaid ettevõteid kõrgemalt hinnata kui mudelil.

Võttes arvesse tulemusi, mis magistritöös väljapakutud tööriista prototüübi poolt saadi, võib öelda, et vastused ja nende kattuvus ning sarnasus jätkusuutlikkuse spetsialistde poolt antud tulemustega indikeerivad, et prototüübi tulemused on head, aga

Tabel 3. Jätkusuutlikkuse spetsialisti vastused võrreldes GPT-4 Turbo vastustega

Ettevõte	**Väiksem**	**Suurem**	**Protsent (%)**
AUGA group	10	0	0.00
EfTEN Real Estate Fund	1	5	**83.33**
Enefit Green	10	18	64.29
Ignitis grupė	10	51	**83.61**
Merko Ehitus	9	12	57.14
Nordecon	10	21	67.74
Pieno žvaigždės	14	4	22.22
TKM Grupp	24	1	4.00
Tallinna Sadam	0	22	**100.00**
Tallinna Vesi	6	7	53.85
Vilkyškių pienin	12	2	14.29

mitte täiuslikud. Selles võib olla, mitu aspekti, näiteks on hindamine subjektiivne ja vastused erinevadki seetõttu. Samuti võib olla, et sisendandmete maht on mudeli jaoks liiga väike ehk spetsialistidel on teavet rohkem kui prototüübile anti. Kindlasti saab tulemuste põhjal öelda, et magistritöös saadud tulemused on lubavad ja järgnevas peatükis väljapakutud edasiarneduse võimalusi rakendades on võimalik saada veelgi täpsemad ja andmetepõhisemad tulemused.

Samuti tuleb märkida, et tänapäeval on ettevõtete jaoks oluline äriline väärtus saada kiiresti ja võimalikult väikeste kuludega. Magistritöös väljapakutud automatiseeritud lahendus annab võimaluse ettevõttel säästa aega ja ressurssi. Kui muidu võtaks sarnase eelanalüüsi tegemine jätkusuutlikkuse spetsialistil aega umbes 4-6 tundi, siis automatiseeritud versiooniga on see võimalik täita minutitega.

6 Tuleviku töö

Magistritöö raames loodud tööriista prototüüp, mis võimaldab jätkusuutlikkuse eelanalüüsi automatiseerimist, on kujunenud oluliseks sammuks finantsasutuste suutlikkuses kiiresti ja tõhusalt hinnata ettevõtete ESG-näitajaid. See prototüüp loodi koostöös finantsasutusega, mis andis võimaluse reaalseks testimiseks ja tagasiside saamiseks. Prototüübi arendus ja selle esialgne testimine on näidanud potentsiaali ESG analüüsi kiirendamiseks ja süvendamiseks. Kuna prototüübi esmane ülesanne oli uurida analüüsi automatiseerimise võimalusi, siis edasine arendustöö ja selle täiustamine toimub finantsasutuse siseselt. Selline lähenemine võimaldab asutusel kohandada tööriista vastavalt oma vajadustele ja integreerida see sügavamalt oma igapäevastesse tööprotsessidesse.

Järgnevates alapeatükkides käsitleb magistritöö autor mitmeid võimalikke suundi, kuidas magistritöö käigus loodud tööriista võiks edasi arendada ja täiustada. Need suunad põhinevad nii magistritöö käigus saadud tagasisidel, kui ka arvestades tehnoloogilisi võimalusi. Esmane samm tööriista täiustamisel on andmehalduse ja töötlusprotsesside optimeerimine. Teiseks, kuna järgnevatel aastatel tekib kohustus ettevõtetel esitada ESG aruanded, siis on sisendandmete mahtu võimalik suurendada valdkonna spetsiifilste andmetega. Kolmandaks aitab kaasa tööriista usadlusväärsemaks ja täpsemaks muutmisel RAG mudeli rakendamine.

6.1 Andmete maht

Magistritöö raames keskenduti esialgsele uurimusele, et hinnata, kas sellise tööriista arendamine võiks andmeid lubavalt analüüsida. Kuna tööriista prototüüpimine hõlmas GPT-4 Turbo API kasutamist, mis on tasuline teenus, tuli arvestada piiratud eelarvega. Tartu Ülikooli poolt jagatud API võti võimaldas teatud arvu päringuid, mille tõttu tuli andmemahu osas teha kompromisse. Magistritöö käigus kasutati väiksemat andmemahtu, mis koosnes peamiselt ESG aruannetest või aastaruannetest saadud eelinfost enne konsolideeritud finantsinfot.

Selline lähenemine võimaldas hinnata tööriista prototüübi esialgset efektiivsust ja analüüsivõimet. Siiski on oluline märkida, et suurema ja spetsiifilisema andmemahu kasutamine võiks märkimisväärselt parandada analüüsi täpsust. Kui ressursid lubaksid, võiks GPT mudelile anda rohkem spetsiifilist sisendinfot, mis võiks sisaldada mitte ainult formaalseid aasta- ja ESG aruandeid, vaid ka ettevõtte kodulehtedel, ajakirjades ja muudes avalikes allikates avaldatud teavet.

Veebist andmete kraapimine ja suurema andmekogumi kasutamine aitaks luua rikkalikuma andmebaasi, mis võimaldaks keelemudelil teha informeeritumaid analüüse ja pakkuda täpsemaid hinnanguid ettevõtete jätkusuutlikkuse kohta. See mitte ainult ei laiendaks analüüsi ulatust, vaid ka suurendaks selle usaldusväärsust, andes võimaluse arvestada mitmekülgsemaid perspektiive ja andmeallikaid.

Tulevikus, ressursside lubades, võiks kaaluda võimalusi API päringute limiidi tõstmiseks või alternatiivsete lahenduste leidmiseks, nagu avatud lähtekoodiga keelemudelid või institutsionaalsed partnerlused, mis võimaldaksid juurdepääsu suuremale hulgale andmetele ilma lisakuludeta. See võimaldaks magistritöö käigus loodud tööriista edasi arendada, integreerides suuremaid andmekogumeid ja pakkudes seeläbi veelgi täpsemaid ja usaldusväärsemaid analüütilisi tulemusi.

6.2 Kestlikkusaruannete regulatsioonid

Magistritöös on kasutatud sisendandmetena Balti börsi põhinimekirjas olevate ettevõete aastaaruandeid. Aastaaruanded on sisendandmeteks suurele keelemudelile GPT-4 Turbo ja sisaldavad domeeni spetsiifilist infot teatud määral kuna aastaaruannetes on kajastatud ka kestlikkuse informatsioon, kuid igal ettevõtel on see varieeruv ja mõningatel puudulik. Järgnevas peatükis toob magistritöö autor välja regulatsioonid, mis hakkavad ettevõtetele kehtima järgnevatel aastatel kestlikkuse aruandlusele. See tähendab, et tulevikus on võimalik anda suurele keelemudelile veelgi rohkem domeeni spetsiifilist sisendinfot.

Kestlikkusaruandluse direktiiv on vastuvõetud Euroopa Liidus 14.detsmbril 2022. aastal, mis sätestab nii kestlikkusaruandluse standardid kui ka ühtse elektroonilise aruandlusvormingu. Ühtne aruandlusvorming on masinloetav, mis aitab kaasa andmetöötlusprotsesside lihtsustamisele. Euroopa riigi ettevõtetel on kohustus vastav direktiiv üle võtta 6. juuliks 2024. aastaks [EU2].

Euroopa Liidu liikmesriikides on sätestatud kohustus esitada kestlikkusaruanded alates 2024. aasta majandusaastast. Esialgu puudutab see suuri avaliku huvi üksusi, kelle töötajate arv ületab 500. Aastast 2025 laieneb see nõue kõigile suurettevõtetele, kellel kaks järjestikust aastat täituvad vähemalt kaks järgmistest tingimustest: vähemalt 250 töötajat, müügitulu vähemalt 50 miljonit eurot või bilansimaht vähemalt 25 miljonit eurot. Alates 2026. aastast kehtib aruandluskohustus ka kõigile ülejäänud börsil noteeritud ettevõtetele, välja arvatud mikroettevõtted, ning samuti väikestele ja mittekeerukatele krediidiasutustele ning kaptiivkindlustusandjatele [Ees24].

6.3 Tagasisidega rikastatud genereerimine

RAG (ingl *Retrieval Augmented Generation*) ehk tagasisidega rikastatud genereerimine on tehnika, mis täiustab domeenispetsiifiliste vestlusrobotite võimekust vastata kasutajate päringutele. RAG kasutab kahte komponenti: päringule vastava konteksti otsimise mudelit ja vastuste genereerimise mudelit, mis põhineb suurel keelemudelil. Konteksti täpsus mõjutab otseselt genereeritud vastuse kvaliteeti [KTKT24].

RAG kontseptsiooni kasutamine võimaldaks magistritöös väljapakutud prototüübi tulemusi parandada, pakkudes täpsemat konteksti iga kasutajapäringu jaoks, mis võimaldab LLM-il genereerida asjakohasemaid ja täpsemaid vastuseid. Sellise lähenemisviisi

kasutamisel magistritöö kontekstis saaks jätkusuutlikkuse spetsialist üle vaadata genereeritud vastused ja anda tagasisidet nende asjakohasuse kohta. See võimaldaks spetsialistil hinnata, kas vastused on paranenud või mitte. Tagasiside põhjal saaks mudelit kohandada, et see vastaks järgnevatele päringutele täpsemalt ja asjakohasemalt, integreerides spetsialisti hinnangud ja täiustades õppimisprotsessi.

7 Kokkuvõte

Jätkusuutlikus on tänapäeval üha tähtsam faktor ettevõtetesse investeerimisotsuste tegemisel. Näiteks on S&P 500 ettevõtted alustanud ESG aruannete avalikustamist, mis oluliselt mõjutavad nende turuväärtust ja investorite otsuseid [AH20] ning samuti ka Balti riikides, kus finantsasutused ja investorid on üha enam hakanud keskenduma sellele, kui jätkusuutlikud on ettevõtted, kuhu nad oma kapitali suunavad [ZB21]. Kuigi käesoleval hetkel on jätkusuutlikkuse aruande täitmine ettevõtetele vabatahtlik, siis järgnevatel aastatel muutub see teatud ettevõtete jaoks ka kohustuslikuks.

Jätkusuutlikkuse aruande täitmise üheks osaks on eelanalüüsi koostamine. Tänasel päeval koostavad seda manuaalselt jätkusuutlikkuse spetsialistid, kes peavad läbi töötama ettevõtete aastaaruannetes oleva info, et täita eelanalüüsi küsimustik. Tegemist on aga mahuka ja ajakuluka tööga. Magistritöö eesmärgiks oligi luua prototüüp tööriist jätkusuutlikkuse eelküsimustiku täitmiseks, mis automatiseerib seda protsessi ning küsimustele vastab GPT-4 Turbo mudel. Arvestades asjaolu, et loodud prototüüp aitab oluliselt kiirendada ja lihtsustada eelanalüüsi koostamist, on magistritööl oluline praktiline väärtus eelkõige finantsasutustele, samuti ka investoritele.

Magistritöö autor on loonud tööriista, mis kogub andmed *https://nasdaqbaltic.com/et* Balti börsi veebilehelt, kust laetakse alla Balti börsi põhinimekirjas olevate ettevõtete aastaaruanded. Tööriista loomiseks rakendas autor andmetele vastavaid PDF ning XHTML failide tekstistamise protsesse ning tekstipuhastus tehnikaid. Vastavad puhastatud kujul tekstilised aastaaruanded on sisendiks GPT-4 Turbo mudelile ning vastavate lähtelausete kasutamisega saadakse vastused jätkusuutlikkuse eelanalüüsi küsimustikule.

Magistritöös tehtud analüüsi põhjal saab järeldada, et mudel annab vastuseid, mis on lähedased spetsialistide poolt antud vastustele. Näiteks 8 ettevõtte puhul 11-st saadi sarnased tulemused jätkusuutlikkuse spetsialistiga, kus veamääraks oli 1 punktiline erinevus. Seega on esimese uurimusküsimuse puhul vastused rahuldavad, kuid mitte täiuslikud. Kuivõrd magistritöös loodud tööriist on prototüüp, mida hiljem arendatakse edasi finantsasutuses, kellega autor koostööd tegi, siis ei olnud ka tulemused täielikult ühtivad valideerimisandmetega. Magistritöös pakutud lahendus on aga kindlasti ajaliselt ning ressursiliselt efektiivsem kui traditsiooniline viis.

Magistritöös sai kinnitust hüpotees, et spetsialistid soovivad enda ettevõtteid tihti paremas valguses näidata kui mudel. Seejuures on jätkusuutlikkuse hindamine üsna subjektiivne ehk ka kaks spetsialisti võivad seda hinnata erinevalt. Seega mudeli edasiarendamisel on võimalik saada andmetepõhisemad tulemused kui hetkel spetsialistide poolt antud vastused. Magistritöös olid otsused tehtud ainult aastaruannetes leiduva info põhjal.

Magistritöö autori loodud prototüüpi on võimalik edasi arendada, muutes tööriista veelgi täpsemaks, usaldusväärsemaks ja efektiivsemaks. Nendeks võiksid olla näiteks andmete mahu suurendamine, kestilikkusaruande regulatsioonid, mis kohustavad ettevõtteid koostama jätkusuutlikkuse aruandeid, mis annab samuti rohkem andmeid, mille alusel

mudel otsuseid saaks teha. Viimaseks pakkus magistritöö autor välja edasiarenduseks RAG ehk tagasisidega rikastatud genereerimise tehnika, mis aitaks mudelit kohandada, et see vastaks järgnevatele päringutele täpsemalt ja asjakohasemalt, integreerides mudelisse spetsialisti hinnangud ja täiustades õppimisprotsessi.

Viidatud kirjandus

[AH20] Bahaaeddin Alareeni and Allam Hamdan. Esg impact on performance of us sp 500-listed firms. *Corporate Governance*, 2020(CG-06-2020-0258):1–18, October 2020.

[AL19] Yrr Ahlklo and Carin Lind. E, s or g? a study of esg score and financial performance. Master of Science Thesis, KTH Industrial Engineering and Management, January 2019. https://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1313629.

[Arm20] Anona Armstrong. Ethics and esg. Victoria University Business School, 2020. https://www.vu.edu.au/business-school.

[AS23] Inc ArtifexŠoftware. Pymupdf: A python library for pdf manipulation. GitHub Repository, 2023. https://github.com/pymupdf/PyMuPDF.

[CNW+24] Zhibo Chu, Shiwen Ni, Zichong Wang, Xi Feng, Chengming Li, Xiping Hu, Ruifeng Xu, Min Yang, and Wenbin Zhang. History, development, and principles of large language models—an introductory survey. ArXiv preprint arXiv:2402.06853, February 2024. https://arxiv.org/abs/2402.06853.

[Ees24] Eesti Finantsinspektsioon. Kestlikkusaruandlus. https://www.fin.ee/finantspoliitika-valissuhted/arvestusvaldkond/kestlikkusaruandlus#kestlikkusaruande-st, 2024. Accessed on 10th April 2024.

[EU2] Euroopa Liidu direktiiv 2022/2464. https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32022L2464&from=ET. Accessed on 10th April 2024.

[Fou23a] PythonŠoftware Foundation. Glob: Unix style pathname pattern expansion. Python Standard Library, 2023. https://docs.python.org/3/library/glob.html.

[Fou23b] PythonŠoftware Foundation. Os: Miscellaneous operating system interfaces. Python Standard Library, 2023. https://docs.python.org/3/library/os.html.

[FW20] Mark L. Frigo and Ray Whittington. Sasb metrics, risk, and sustainability. Strategic Finance, April 2020. https://www.imanet.org/strategic-finance.

[JW] Sam Hays Michael Sandborn Carlos Olea Henry Gilbert Ashraf Elnashar Jesse Spencer-Smith Douglas C.Šchmidt Jules White, Quchen Fu. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. Department of Computer Science, Vanderbilt University, Tennessee Nashville, TN, USA.

[KTKT24] Mandar Kulkarni, Praveen Tangarajan, Kyung Kim, and Anusua Trivedi. Reinforcement learning for optimizing rag for domain chatbots. In *Proceedings of the AAAI 2024 Workshop on Synergy of Reinforcement Learning and Large Language Models*, Seattle, Washington, USA, 2024. Association for the Advancement of Artificial Intelligence (www.aaai.org).

[MMN+24] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. ArXiv preprint arXiv:2402.06196, February 2024. https://arxiv.org/abs/2402.06196.

[Ope22] OpenAI. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt, 2022. Accessed on 10th April 2024.

[PdG22] Alain Pietrancosta and Alexis Marraud des Grottes. Trends - what the boards of all companies should know about esg regulatory trends in europe. Sorbonne Law School at the University of Paris, August 2022. https://ssrn.com/abstract=4206521.

[PHS+22] Lucy Pérez, Vivian Hunt, Hamid Samandari, Robin Nuttall, and Krysta Biniek. Does esg really matter—and why? McKinsey Company, August 2022. https://www.mckinsey.com/.

[Ric23] Leonard Richardson. Beautiful soup: A library for pulling data out of html and xml files. GitHub Repository, 2023. https://www.crummy.com/software/BeautifulSoup/.

[SAS] Sasb standards. https://sasb.ifrs.org/standards/. Accessed: 2024-04-16.

[SSM+23] Kimya Khakzad Shahandashti, Mithila Sivakumar, Mohammad Mahdi Mohajer, Alvine B. Belle, Song Wang, and Timothy C. Lethbridge. Evaluating the effectiveness of gpt-4 turbo in creating defeaters for assurance cases. York University and University of Ottawa, January 2023. https://www.yorku.ca/ and https://www.uottawa.ca/.

[ZB21] Ilze Zumente and Jūlija Bistrova. Do baltic investors care about environmental, social and governance (esg)? *Entrepreneurship and Sustainability Issues*, 8(4):349–362, June 2021. Accessed: 2024-04-16.

[ZSC+23] YiŽou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, HongXiang Tong, Lei Xiao, and Wenwen Zhou. Esg llm: Leveraging large language models for environmental, social, and governance data extraction. Alibaba Cloud, Hangzhou, China; Department of Earth System Science, Tsinghua University, Beijing, China; Department of Environmental Science and Engineering, Sun Yat-Sen University, Guangzhou, China, December 2023. https://arxiv.org/abs/2312.17264.

Lisad

I. GitHub repositooriumi link

https://github.com/BrandonLoorits/Loorits_Andmeteadus_2024.git

IV. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Brandon Loorits**,
*(*autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   Ettevõtete jätkusuutlikkuse eelanalüüs aasta aruannete põhjal Balti börsi ettevõtete näitel,
   *(*lõputöö pealkiri)

   mille juhendajad on Mark Fišel ja Lehar Oha,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Brandon Loorits
15.05.2024