TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Mihkel Järviste

Sisupõhiste ja semantiliste vektorkujutustest hübriidmudelite võrdlus e-kaubanduse soovitussüsteemides

Bakalaureusetöö (9 EAP)

Juhendajad:
Andres Järviste, MSc
Margus Niitsoo, PhD

Tartu 2025

Sisupõhiste ja semantiliste vektorkujutustest hübriidmudelite võrdlus e-kaubanduse soovitussüsteemides

Lühikokkuvõte:

Käesoleva bakalaureusetöö eesmärk oli võrrelda kahte hübriidsoovitussüsteemi Steam videomängude andmestikul. Hübriidsüsteemid on traditsioonilisel sisupõhisel TF-IDF meetodil ning kaasaegsemal semantilisel SBERT lausevektoritel põhinev. Mõlemad lähenemised ühendati mängude üldise hinnanguga. Süsteeme hinnati testkorpusel, kasutades mängude mitmekülgsetel tunnustel (žanrid, kategooriad, kasutajate tag'id) ja Jaccardi indeksil põhinevat tõese sarnasuse (GT) definitsiooni ning standardseid jõudlusnäitajaid.

Uurimuse tulemused näitasid järjepidevalt, et sisupõhine (TF-IDF) hübriidmudel edestas semantilist (SBERT) hübriidmudelit. Järeldati, et sisupõhise mudeli paremus tulenes tõenäoliselt GT olemusest, mis premeeris selgesõnaliste tunnuste kattuvust, ning Steam mängude andmestiku spetsiifikast, kus konkreetsed tunnused on sarnasuse hindamisel olulised. Töö rõhutab, et soovitussüsteemide efektiivsus sõltub kontekstist ning traditsioonilised meetodid võivad teatud tingimustel anda paremaid tulemusi.

Võtmesõnad:
Soovitussüsteem, SBERT, sõnade vektoresitused

CERCS:
P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine
P176 Tehisintellekt

Comparison of Hybrid Models Based on Content and Semantic Vector Representations in E-commerce Recommender Systems

Abstract:

This bachelor's thesis aimed to compare two hybrid recommender systems on a Steam video game dataset. The hybrid systems are based on traditional content-based TF-IDF similarity and modern semantic SBERT sentence embeddings. Both approaches were combined with overall game ratings. The systems were evaluated on a test set using a ground truth (GT)

defined by the overlap of diverse game features (genres, categories, user tags) via the Jaccard index, employing standard performance indicators.

The results consistently demonstrated that the content-based (TF-IDF) hybrid model outperformed the semantic (SBERT) hybrid model. It was concluded that the superiority of the content-based model likely stemmed from the GT's nature, which rewarded explicit feature overlap, and the specifics of the Steam games dataset, where distinct features are crucial for similarity assessment. The thesis highlights that recommender system effectiveness is context-dependent, and traditional methods can yield superior results under certain conditions.

Visual Abstract:

Keywords:
Recommender system, SBERT, word embedding

CERCS:
P170 Computer science, numerical analysis, systems, control
P176 Artificial intelligence

Sisukord

Sissejuhatus

Soovitussüsteemid e-kaubanduses on viimastel aastatel kiiresti kasvanud ja muutunud väga mõjukaks valdkonnaks. Suur konkurents nõuab, et ettevõtted suudaksid klientidele pakkuda kiiret ja isikupärastatud teenust. Soovitusüsteemid on muutunud üha olulisemaks, kuna inimesed seisavad silmitsi info üleküllusega. Sellised süsteemid võimaldavad kasutajatel tõhusamalt vajalikku teavet leida, pakkudes neile just seda informatsiooni, mis on neile kõige asjakohasem [1].

Selle eesmärgi saavutamiseks on välja töötatud mitmeid lähenemisviise. Klassikaliselt jaotatakse soovitussüsteemid kolme põhikategooriasse: sisupõhised, koostööpõhised ja hübriidsüsteemid [2]. Sisupõhised süsteemid analüüsivad esemete omadusi ning püüavad leida kasutajale sarnase sisuga tooteid, lähtudes tema varasemast eelistuste profiilist [2]. Koostööpõhised süsteemid seevastu arvestavad teiste kasutajate hinnanguid või käitumist. Eeldusel, et kahel kasutajatel on minevikus olnud sarnased eelistused, siis ühe kasutaja lemmikud sobivad soovitada ka teisele [2]. Hübriidsed süsteemid ühendavad erinevaid lähenemisviise eesmärgiga ületada ühe meetodi kitsaskohad teise tugevaid külgi kasutades [2]. Igal lähenemisel on oma tugevused ja puudused, mida käsitletakse täpsemalt töös Teoreetilise tausta peatükis.

Käesoleva töö eesmärk on võrrelda kahte hübriidset soovitussüsteemi ja nende tulemusi. Esiteks vaadeldakse traditsioonilist hübriidset lahendust, mis ühendab sisupõhist ja koostööpõhist filtrit. Teiseks analüüsitakse uuenduslikumat hübriidlahendust, mis kombineerib koostööpõhise meetodi ja semantilise filtreerimise, kasutades mängude kirjelduse analüüsimiseks lausevektoreid. Mõlemaid süsteeme rakendatakse samale andmestikule, et võimaldada ausat võrdlust.

1. Teoreetiline taust

Traditsioonilised soovitussüsteemide loomise meetodid tuginevad kasutajate hinnangutele või toodete sisulistele omadustele, kuid tavaliselt piirduvad hinnangutele põhineva käitumisanalüüsiga või sõnapõhise tekstianalüüsiga. Hübriidlahendused ühendavad need kaks lähenemist, et tasakaalustada täpsust ja mitmekesisust. Kaasaegsed semantilised mudelid ja suured keelemudelid avavad võimaluse tekstist sügavama tähendusliku teabe äratundmiseks. Järgnevad alapeatükid pakuvad põhjalikumat ülevaadet iga lähenemisviisi põhimõtetest, rakendustest ning nendega seotud eeliseid ja väljakutseid.

1.1 Ülevaade soovitussüsteemidest

Soovitussüsteemid on tarkvaralahendused, mille eesmärk on pakkuda kasutajatele isikustatud soovitusi, lähtudes nende varasematest eelistustest, käitumismustritest ja huvidest. Viimaste aastakümnete jooksul on soovitussüsteemid muutunud digikeskkondades asendamatuks, aidates kasutajatel orienteeruda järjest suurenevates info- ja tootekogustes. Traditsiooniliselt eristatakse soovitussüsteemides kolme peamist lähenemisviisi: koostööpõhised süsteemid (collaborative filtering), sisupõhised süsteemid (*content-based filtering*) ning nende kahe kombinatsioonina loodud hübriidsüsteemid [3] [4].

Koostööpõhised soovitussüsteemid (*collaborative filtering*) tuginevad kasutajate varasematele tegevustele, näiteks ostudele või toodetele antud hinnangutele. Sellised süsteemid eeldavad, et kui teised kasutajad on varem näidanud sarnast käitumist või eelistanud sarnaseid tooteid, siis tõenäoliselt kattuvad nende huvid praegu süsteemi kasutava inimese omadega. Peamised meetodid, mida kasutatakse koostööpõhistes süsteemides on kasutaja- või tootepõhised lähenemised, mis võrdlevad kasutajate või toodete sarnasusi ja pakuvad soovitusi vastavalt teiste kasutajate või sarnaste toodete alusel saadud informatsioonile [4].

Sisupõhised soovitussüsteemid lähtuvad seevastu toodete omadustest ja kasutajate eelnevast huvist nende omaduste suhtes. Sellisel juhul loob süsteem iga kasutaja jaoks unikaalse profiili, mis põhineb toodete omadustel, mida kasutaja on varasemalt eelistanud. Sisupõhistes soovitustes kasutatakse sageli toodete kirjeldusi, märksõnu ja muid sisulisi atribuute, mille põhjal leitakse sarnaseid tooteid kasutaja juba ostetud või vaadatud toodetega. Selle meetodi üheks tugevuseks on võime pakkuda soovitusi ka olukorras, kus toote kohta puuduvad veel teiste kasutajate hinnangud või eelnev ajalugu [3].

Hübriidsoovitussüsteemid ühendavad eelnevaid lähenemisi, püüdes sellega kompenseerida mõlema meetodi puudujääke. Tavaliselt ühendatakse koostöö- ja sisupõhised meetodid viisil, mis võimaldab tasakaalustada nende tugevusi ja vähendada puudusi. Selliste süsteemide eelised tulevad selgelt esile e-kaubanduses ja teistes valdkondades, kus oluline on pakkuda soovitusi nii pikaajalistele kui ka uutele kasutajatele, samal ajal tagades soovituste piisava mitmekesisuse ja kvaliteedi [3] [4].

Viimastel aastatel on soovitussüsteemide valdkonnas tehtud suuri edusamme tänu tehisintellekti ja süvaõppe meetodite kasutusele võtmisele. Näiteks võimaldab närvivõrkudel põhinev lähenemine ühendada erinevat tüüpi andmeid – nagu tekstid, pildid, helid ja kasutaja käitumise mustrid – ühtsesse terviklikku mudelisse. Sellised uued AI-põhised süsteemid pakuvad traditsiooniliste meetoditega võrreldes suuremat paindlikkust ja täpsust, olles seeläbi muutumas järjest populaarsemaks valikuks soovitussüsteemide rakendamisel [5][7].

1.2 Soovitussüsteemide olulisus e-kaubanduses

Soovitussüsteemid on kujunenud e-kaubanduse lahutamatuks osaks, kuna nende abil on võimalik märkimisväärselt parandada kasutajate kogemust, suurendada klientide rahulolu ja tõsta ettevõtete majanduslikku tulemust. Kuna digitaalsete toodete ja teenuste hulk pidevalt kasvab, muutub kasutajatel sobivate valikute iseseisvalt tegemine järjest keerulisemaks. Sellises olukorras muutub kvaliteetsete soovituste pakkumine järjest kriitilisemaks konkurentsieeliseks [3].

Üheks eredaimaks näiteks e-kaubanduse soovitussüsteemide mõjust on Amazon, kus hinnanguliselt moodustavad soovituste abil tehtud ostud umbes 35% ettevõtte kogumüügist [3]. Soovitussüsteemide mõju on eriti märgatav ka digitaalsete meelelahutusplatvormide puhul, nagu Netflix ja Spotify, kus kasutajale isikustatud sisu pakkumine on otseselt seotud kasutajate püsimisega platvormil ja nende pikaajalise rahuloluga [3].

E-kaubanduses kasutatakse soovitussüsteeme väga mitmekesiselt sõltuvalt konkreetsest valdkonnast ja eesmärgist. Näiteks jaemüügiplatvormidel aitavad soovitussüsteemid suurendada rist- ja lisamüüki, pakkudes ostjatele lisatooteid, mis sobivad kokku nende varasemate ostudega. Sellisel juhul võivad soovitused põhineda nii teiste kasutajate ostukäitumisel kui ka toodete sarnastel omadustel. Reisi- ja hotellibroneerimise platvormidel on soovituste ülesandeks suurendada kasutaja otsustusmugavust ja pakkuda neile just selliseid majutusvõimalusi ja reisielamusi, mis kõige paremini vastavad nende eelistustele ja varasemale käitumisele [3] [4].

Lisaks otsesele majanduslikule kasule aitavad soovitussüsteemid tõsta ka kasutajakogemuse kvaliteeti ja mugavust, pakkudes igaühele individuaalselt kohandatud valikuid ja võimaldades vältida üleliigse info esitamist. Samuti aitavad soovitused leevendada info ülekülluse probleemi ning suurendada vähem tuntud või nišitoote nähtavust, võimaldades kasutajatel avastada neile täiesti uusi ja huvipakkuvaid võimalusi, mille leidmine ilma soovitusteta oleks oluliselt keerulisem [3][5].

Seetõttu on kvaliteetsete ja efektiivsete soovitussüsteemide rakendamine muutunud paljude e-kaubandusettevõtete jaoks kriitiliseks prioriteediks, aidates neil saavutada konkurentsieelist, kasvatada klientide lojaalsust ning parandada pikaajalist äriedu ja jätkusuutlikkust.

1.3 Hinnangu-põhised soovitussüsteemid

Traditsiooniliselt tuginesid soovitussüsteemid peamiselt kasutajate antud hinnangutele või kaudsele tagasisidele, nagu näiteks toodete vaatamine või ostmine. Viimaste aastate jooksul on aga üha rohkem hakatud tähelepanu pöörama kasutajate kirjutatud arvustustele ehk tekstipõhisele tagasisidele. Selliste arvustuste kasutamisel on võimalik saada palju detailsem ülevaade kasutajate eelistustest, toote omadustest ja üldisest kasutajakogemusest, mida pelgalt numbrilised hinnangud ei kajasta [4].

Arvustustel põhinevate soovitussüsteemide keskseks ideeks on see, et kasutajate kirjutatud tekstides sisaldub väärtuslik info, mis aitab mõista mitte ainult seda, kas kasutajale mingi toode meeldib, vaid ka seda, miks see talle meeldib või ei meeldi. Selle informatsiooni eraldamiseks ja kasutamiseks kasutatakse tavaliselt loomuliku keele töötlemise (natural language processing – NLP) meetodeid, näiteks teemade modelleerimist ja aspektipõhist sentimentanalüüsi [6]. Üks tuntumaid lähenemisi on LDA (*Latent Dirichlet Allocation*), mille abil tuvastatakse arvustustest olulised teemad ja nende kohta väljendatud arvamused.

Lisaks traditsioonilistele NLP-meetoditele on viimasel ajal hakatud arvustuste töötlemisel üha rohkem kasutama ka süvaõppe mudeleid, näiteks DeepCoNN (*Deep Cooperative Neural Networks*) ja tähelepanumehhanismidel põhinevaid lahendusi nagu NARRE (*Neural Attentional Rating Regression*). Sellised mudelid võimaldavad paremini eristada olulist infot mürast ning tuua esile just need arvustused või arvustuste osad, mis annavad kasutajate eelistustest kõige täpsema ülevaate [8][9].

Arvustustel põhinevate süsteemide üheks olulisemaks eeliseks on võime vähendada andmete hõreduse probleemi. Kuna isegi vähesed arvustused võivad sisaldada piisavalt infot kasutajate eelistuste kohta, võimaldab see pakkuda täpsemaid soovitusi ka neile kasutajatele või toodetele, mille kohta on kogutud vähe hinnanguid. Lisaks suurendab arvustuste kasutamine süsteemi läbipaistvust ja selgitatavust, kuna soovitusi on võimalik põhjendada konkreetsete kasutajate poolt välja toodud omadustega. Näiteks saab süsteem põhjendada soovitust viitega, et „seda toodet soovitame teile, kuna paljud kasutajad on välja toonud selle vastupidava aku ja kerge kaalu“.

Samas kaasnevad arvustustepõhiste süsteemidega ka teatud väljakutsed. Suurimaks probleemiks on tekstide ebaühtlane kvaliteet ning sisuline ja vormiline varieeruvus, mistõttu võib olla keeruline eristada olulist ja tõsiseltvõetavat infot müra ja kõrvaliste kommentaaride hulgast. Lisaks on tekstianalüüs arvutuslikult kulukas, eriti kui tegemist on süvaõppemudelitega, mistõttu võib tekkida probleeme süsteemi skaleerimisel suurtes andmekogudes [4][6]. Sellegipoolest näitavad viimased uuringud, et arvustuste kaasamine suurendab soovitussüsteemide täpsust ja aitab paremini rahuldada kasutajate vajadusi, pakkudes ühtlasi võimalust selgitada soovituste tagamaid.

1.4 Traditsioonilised hübriidsed soovitussüsteemid

Traditsioonilised hübriidsed soovitussüsteemid on loodud eesmärgiga ühendada kahe või enama erineva soovitusmeetodi tugevused ja tasakaalustada nende puudusi. Enamasti kombineeritakse nendes süsteemides koostööl põhinevaid (CF) ja sisupõhiseid (CBF) soovitusalgoritme, et tagada paremaid tulemusi kui kumbki meetod üksi võimaldaks [3].

Hübriidsüsteemide kasutamine võimaldab leevendada näiteks koostööpõhiste meetodite puudusi, nagu andmete hõredus ja uute toodete või kasutajatega seotud külmkäivituse (*cold-start*) probleemid. Samal ajal aitavad koostööpõhised meetodid vältida sisupõhiste süsteemide tendentsi soovitada liiga sarnaseid tooteid ja suurendada soovituste mitmekesisust ning kasutajate avastamisvõimalusi [3] [4].

Traditsioonilised hübriidsüsteemid kasutavad erinevaid strateegiaid nende meetodite ühendamiseks. Levinud lahenduseks on näiteks kaalutud kombinatsioon, mille puhul antakse erinevatele meetoditele kaalud vastavalt nende eelnevalt tõestatud täpsusele või vastavalt kasutajate või toodete andmete iseloomule. Alternatiivina kasutatakse meetodite vahel dünaamilist ümberlülitamist sõltuvalt olukorrast – näiteks uue kasutaja puhul tugineb süsteem rohkem sisupõhistele soovitustele, samal ajal kui pika ajalooga kasutajate puhul kasutatakse

rohkem koostööpõhiseid meetodeid. Veel üheks levinud lahenduseks on järjestikune ehk kaskaadne lähenemine, kus näiteks sisupõhine soovitusalgoritm genereerib esmase soovitusnimekirja, mida koostööpõhine meetod seejärel täpsustab ja järjestab lõpliku soovituste nimekirja saamiseks [3].

Hübriidsüsteemide kasutamine on laialt levinud mitmesugustes e-kaubanduse valdkondades. Näiteks Netflixi soovitussüsteem kasutab hübriidset lähenemist, ühendades kasutajate vaatamisajaloo põhjal koostööpõhise analüüsi filmide sisuliste omadustega, nagu žanr või näitlejad. Samamoodi kasutatakse rõivaste ja moe e-poodides sageli hübriidseid süsteeme, kus koostööpõhised meetodid kombineeritakse piltide põhjal saadud visuaalsete omadustega, et pakkuda kasutajatele stiililiselt ja visuaalselt sobivaid soovitusi [3].

Hübriidsete süsteemide peamisteks väljakutseteks on nende tehniline keerukus ja ressursside nõudlikkus, eriti kui süsteemid peavad toimima reaalajas ja suurtel andmehulkadel. Lisaks võib olla keeruline leida optimaalset tasakaalu erinevate meetodite vahel ning tagada, et iga kasutatud meetod oleks efektiivselt kaasatud. Samas on arvukad uuringud näidanud, et hoolimata keerukusest annavad hübriidmeetodid oluliselt paremaid tulemusi nii soovituste täpsuse, mitmekesisuse kui ka süsteemi üldise kasutatavuse seisukohalt [3] [4].

1.5 Semantilised soovitussüsteemid ja suurte keelemudelite (LLM) integreerimine

Semantilised soovitussüsteemid tuginevad kasutajate ja toodete vaheliste seoste mõistmisele ning annavad soovitusi, mis põhinevad mitte ainult kasutaja eelneval käitumisel, vaid ka toodete ja kasutajate vaheliste tähenduslike suhete analüüsil. Sellistes süsteemides kasutatakse sageli semantilisi teadmisi sisaldavaid struktuure, nagu näiteks teadmistegraafid (*knowledge graphs*), mis võimaldavad tuvastada sügavamaid ja keerukamaid seoseid toodete vahel [10]. Näiteks võib teadmistegraaf aidata soovitada kasutajale raamatut, mis kuulub samasse žanrisse või on seotud sarnase teemaga nagu tema varem loetud raamatud, isegi kui puuduvad otsesed hinnangute põhjal tuvastatavad seosed.

Lisaks teadmistegraafidele on viimasel ajal hakatud üha enam integreerima soovitussüsteemidesse suuri keelemudeleid (large language models ehk LLM-id), nagu OpenAI GPT-3 või GPT-4. Need mudelid on treenitud ulatuslikel tekstikorpustel ning need suudavad seetõttu mõista loomulikku keelt ja luua semantilisi representatsioone, mis võimaldavad oluliselt täpsemalt mõista kasutajate huvisid ja toodete omadusi [7]. Näiteks

võib LLM-i abil saadud tekstirepresentatsioon aidata ühendada kasutaja arvustustest saadud tagasiside toodete kirjeldustega isegi siis, kui need tekstid ei kasuta samu märksõnu või väljendeid.

Lisaks täpsemale semantilisele mõistmisele võimaldavad LLM-id luua ka dialoogipõhiseid ehk vestluslikke soovitussüsteeme. Sellised süsteemid ei piirdu pelgalt toodete soovitamisega, vaid suudavad pidada kasutajatega loomulikku vestlust, selgitades soovituste põhjuseid ja kohanda soovitusi vastavalt kasutajapoolsele tagasisidele reaalajas [13]. Sellise interaktiivse süsteemi kasutamine võib oluliselt suurendada kasutajate usaldust ja rahulolu, kuna süsteem on võimeline selgelt põhjendama oma soovitusi ja täpsustama kasutaja eelistusi interaktiivse dialoogi käigus.

Selliste süsteemide praktiliseks realiseerimiseks kasutatakse sageli arendusraamistikke, näiteks LangChain, mis võimaldab ühendada suuri keelemudeleid olemasolevate andmebaaside ja teiste rakendustega. Sellise integratsiooni korral ei pea süsteem ise kogu teavet omama, vaid saab päringute kaudu hankida vajaliku teabe muudest andmeallikatest, vähendades nii võimalust, et LLM soovitusi „välja mõtleb“ ehk halutsineerib [14]. See lahendus võimaldab ühendada keelemudelite paindlikkuse ja loomingulisuse traditsiooniliste soovitusmudelite usaldusväärsuse ja täpsusega.

Kuigi suurte keelemudelite kasutamisel on palju eeliseid, nagu suurem semantiline täpsus ja parem selgitatavus, kaasnevad nende kasutamisega ka mõned väljakutsed. Peamisteks probleemideks on LLM-ide suur arvutuslik nõudlikkus ja kõrged kulud, mis võivad piirata nende praktilist rakendatavust suuremahulistes süsteemides. Samuti on väljakutseks tagada LLM-ide soovituste stabiilsus ja järjepidevus, kuna väikesed muudatused sisendis võivad põhjustada ootamatuid erinevusi väljundis [7]. Seetõttu kasutatakse LLM-e sageli kombineeritult traditsiooniliste süsteemidega, et saavutada tasakaal innovaatilise semantilise analüüsi ja praktilise usaldusväärsuse vahel.

1.6 Soovitussüsteemide väljakutsed ja piirangud

Vaatamata sellele, et soovitussüsteemid on viimase kümnendi jooksul märkimisväärselt arenenud, seisavad nii traditsioonilised kui ka tehisintellektil (AI) põhinevad lahendused silmitsi mitmete püsivate väljakutsete ja piirangutega. Üheks suurimaks probleemiks traditsiooniliste koostööpõhiste (CF) süsteemide puhul on andmete hõredus ja külmkäivituse probleem. Kuna paljud kasutajad hindavad vaid väheseid tooteid ning uutel toodetel või kasutajatel puudub vajalik ajalugu, ei suuda traditsioonilised meetodid alati alguses täpseid

soovitusi anda [4]. Ehkki sisupõhised (CBF) ja hübriidsed lahendused aitavad seda probleemi osaliselt leevendada, jääb see siiski oluliseks väljakutseks e-kaubanduse kontekstis.

Teiseks traditsiooniliste meetodite piiranguks on nende vähene võime mõista, miks kasutajatele mingid tooted meeldivad või ei meeldi. Need meetodid põhinevad sageli üksnes varasemate hinnangute mustritel ning ei analüüsi põhjalikult toodete omadusi ega kasutajate konkreetseid eelistusi. Seetõttu võivad traditsioonilised süsteemid soovitada tooteid, mis on küll statistiliselt sarnased varem ostetud kaupadele, kuid ei vasta täielikult kasutaja tegelikele eelistustele ega vajadustele [3].

AI-põhised semantilised süsteemid, sealhulgas suurtel keelemudelitel (LLM) põhinevad soovitajad, pakuvad küll märkimisväärseid eeliseid semantilise arusaamise ja soovituste selgitatavuse osas, kuid ka nendega kaasnevad uued probleemid. Üheks suurimaks väljakutseks on LLM-ide kõrge arvutuslik keerukus ja nendega seotud kulud. Suurte mudelite kasutamine reaalajas soovitussüsteemides võib muutuda liiga aeglaseks või kulukaks, mistõttu on vajalik hoolikalt planeerida nende integreerimist traditsiooniliste süsteemidega, tagamaks tasakaal jõudluse ja kulude vahel [7].

Lisaks võivad LLM-põhised süsteemid olla vähem stabiilsed kui traditsioonilised soovitusalgoritmid. Väikesed erinevused sisendtekstis või küsimuste sõnastuses võivad tekitada ootamatuid ja soovimatuid erinevusi väljundites, mis raskendab selliste süsteemide järjepidevat kasutamist ja hindamist. Lisaks tuleb arvestada võimalusega, et keelemudelid võivad ilma piisava kontrollita soovitada tooteid, mis reaalsuses ei eksisteeri või ei vasta kasutajate soovidele, mis kahjustab süsteemi usaldusväärsust ja kasutajakogemust [14].

Samuti võib olla keeruline mõõta ja hinnata uute AI-põhiste süsteemide tegelikku tõhusust ja väärtust, sest traditsioonilised mõõdikud (näiteks täpsus ja tagasikutsumine) ei pruugi kajastada täielikult kasutajate tegelikku rahulolu ega süsteemi võimet soovitusi hästi põhjendada. Selleks on vaja välja töötada uusi hindamismeetodeid ja mõõdikuid, mis suudaksid objektiivselt hinnata nii soovituste kvaliteeti kui ka nende selgituste usaldusväärsust ja mõju kasutajate rahulolule [7].

Kokkuvõttes seisavad nii traditsioonilised kui ka AI-põhised soovitussüsteemid silmitsi oluliste piirangutega, mille edukas ületamine nõuab läbimõeldud süsteemide disaini ja nutikate tehniliste lahenduste kasutamist. Sageli on kõige tõhusamaks lahenduseks just nende kahe lähenemise hübriidne kombineerimine, mille käigus ühendatakse traditsiooniliste meetodite stabiilsus ja tõhusus AI-lahenduste innovaatilisuse ja paindlikkusega.

1.7 Soovitussüsteemide tulevikutrendid ja arengusuunad

Tulevikus on oodata, et soovitussüsteemid muutuvad veelgi interaktiivsemaks, selgitatavamaks ja kasutajakesksemaks. Üks olulisemaid arengutrende on dialoogipõhiste ehk vestluslike soovitussüsteemide laiem kasutuselevõtt. Sellised süsteemid ei piirdu pelgalt ühekordsete soovitustega, vaid suhtlevad kasutajatega loomuliku keele vahendusel, kohandades soovitusi dünaamiliselt vastavalt kasutajate reaalajas antud tagasisidele (Jannach jt., 2021). Selline lähenemine suurendab kasutajate kaasatust ja võimaldab pakkuda paremini personaliseeritud soovitusi, kuna süsteem saab kasutajalt pidevalt täiendavaid signaale tema huvide ja eelistuste kohta.

Teine oluline tulevikusuund on soovituste selgitatavuse ja läbipaistvuse suurendamine. Kasutajad soovivad üha rohkem mõista, miks neile just teatud tooteid soovitatakse, ning ootavad süsteemidelt selgeid ja usaldusväärseid põhjendusi. Semantilised soovitussüsteemid, eriti need, mis kasutavad suuri keelemudeleid ja teadmistegraafe, on võimelised pakkuma detailseid ja sisukaid selgitusi, aidates sel viisil suurendada kasutajate usaldust süsteemi vastu [14]). Samuti pööratakse tulevikus rohkem tähelepanu süsteemide läbipaistvusele ja nende eetikale, et vältida soovituste kallutatust ja tagada kasutajate õiglane kohtlemine.

Kokkuvõttes liiguvad soovitussüsteemid järjest nutikamate, interaktiivsemate ja kasutajakesksemate lahenduste poole, mis mitte ainult ei paku täpseid soovitusi, vaid tagavad ka süsteemi usaldusväärsuse, läbipaistvuse ning kasutajate privaatsuse ja eetilise kohtlemise.

2. Hübriidsüsteemide loomine

Eesmärgiks on luua kaks mudelit, mida hiljem omavahel võrrelda: esimene põhineb traditsioonilisemal sisupõhisel sarnasusel (TF-IDF) kombineerituna mängude üldise hinnanguga, ning teine kasutab sarnasuse leidmiseks kaasaegsemat semantilist lähenemist (SBERT lausevektorid), samuti kombineerituna mängude üldise hinnanguga [17][18]. Alustatakse andmestiku valiku ja eeltöötlusega, millele järgneb mängude jaoks vajalike tunnuste (tekstilised sisendvektorid, kaalutud hinnangud) ja sarnasusmeetodite kirjeldus.

2.1 Andmestiku kogumine

Käesolev uurimustöö tugineb avalikult kättesaadavale Steam mängude andmestikule, mille autoriteks on Artemiy Ermilov ja kaasautorid ning mis on avaldatud Kaggle'i platvormil. Andmestik on loodud kombineerides Steami poelehtedelt andmete automaatset kraapimist (*scraping*) ning kasutades Steam API ja Steam Spy teenuseid täiendava informatsiooni kogumiseks. Esialgsel kujul sisaldas see andmestik üle 90 000 mängukirje. Kuna kirjed koguti ja uuendati ühe hetktõmmisena, peegeldavad need platvormi seisu märts 2025 seisuga.

Andmestik valiti just seetõttu, et sisaldas mitut tüüpi teavet ning võimaldab rakendada nii sisupõhiseid kui koostööpõhiseid soovitusmeetodeid. Tekstiväljad pakuvad loomuliku keele materjali, mille abil hinnata mängude sisulist lähedust. Arvustused annavad võimaluse teha järeldusi mängude populaarsuse ja kasutajarahulolu kohta.

2.2 Eeltöötlus ja valimi moodustamine

Esialgne ligi 90 000 mängukirjet sisaldav andmekogu läbis mitmeastmelise eeltöötlus- ja filtreerimisprotsessi, et moodustada soovitussüsteemide loomiseks ja testimiseks sobiv kvaliteetne valim. Protsessi eesmärk oli eemaldada ebapiisava informatsiooniga või vähelevinud mängud ning tagada töödeldavate andmete ühtlus.

2.2.1 Tekstilise sisu kvaliteet

Valim lähtus juba varasemalt puhastatud andmebaasist, millest olid eemaldatud HTML-sildid ja muud vormindusjäägid. Täiendavalt rakendati kaks sammu. Kõigepealt jäeti välja need kirjed, mille detailne kirjeldus sisaldas vähem kui 35 sõna, sest nii lühikesed kirjeldused ei paku mudelile piisavat semantilist konteksti. Seejärel normaliseeriti allesjäänud tekstid: kõik väljad sunniti tekstitüübiks, tühjad väärtused asendati tühja sõnega ning järjestikused tühikud

ja reavahetused koondati üheks tühikuks. Need sammud olid vajalikud, et saavutada kombineeritud andmeväli, mida saame kasutada mõlema hübriidsüsteemi loomiseks.

2.2.2 Populaarsuse miinimumnõue

Andmestikule seati miinimumnõue, et iga valimisse jääv mäng oleks kogunud piisavalt kasutajate arvustusi ning seeläbi pakuks usaldusväärset populaarsussignaali. Läveks valiti 313 arvustust, mis vastab alumise veerandi piirile esialgses andmekogumis. Kirjed, mis sellest mahust allapoole jäid, eemaldati. Puuduolevad numbrilised väärtused veergudes *estimated_owners*, *peak_ccu* ja *num_reviews_total* asendati nullidega, et vältida hilisema analüüsi katkestamist. Seejärel loodi kombineeritud populaarsusskoor, mis prioritiseeris hinnangulist omanike arvu, seejärel tipptunni samaaegsete mängijate arvu ja viimasena arvustuste koguarvu. Seega eelistati hindamiseks pigem mängijate koguarvu, mitte tipptunni populaarsust.

2.2.3 Keeletuvastus

Järgmise sammuna teostati keeletuvastus. Kirjelduste keel tuvastati Python-i teegiga langdetect. Keeletuvastus rakendati esmalt 7000 populaarseimale mängule, et vähendada arvutuskoormust. Iga mängu detailne kirjeldus edastati funktsioonile, mis tagastas keelekoodi. Kui keelekood ei viidanud inglise keelele, tuvastus ebaõnnestus. Analüüsi jätkamiseks jäeti alles üksnes need mängud, mille kirjeldus tuvastati inglise keelena. Niiviisi tagati, et järgmistes etappides kasutatavad tekstipõhised mudelid töötlevad semantiliselt ühekeelset korpust.

2.2.4 Lõplik valim

Rakendatud tekstilise kvaliteedi, populaarsuse ja keeletuvastuse filtrite järel hinnati ülejäänud mänge varasemalt punktis 2.2.2 kirjeldatud kombineeritud populaarsusskoori alusel. Mängud järjestati skoori alusel kahanevas järjekorr 45as ning analüüsiks valiti 5000 kõrgeima skooriga pealkirja. See lähenemine tagas, et valikus on piisavalt laialt levinud ja rikkalikult tagasisidet saanud mänge, võimaldades nii koostöö- kui sisupõhiste mudelite statistiliselt usaldusväärset hindamist.

2.3 Tekstväljade ühendamine

Pärast lõplikku valimit koostati iga mängu kohta üks koondtekst, mis koondas kõik sisulised kirjeldusväljad. Selleks liideti lühikirjeldus ja detailne kirjeldus, millele lisati järjest žanrid,

kategooriad ning platvormi kasutajate määratud märksõnad. Kirjeldusväljade koondtekst salvestati väljale *combined_text*. Märksõnade puhul kasutati ainult nende nimetusi, eemaldades esialgse informatsiooni sageduse kohta, et tagada ühtlane tekstikuju. See ühtne tekstiväli oli aluseks nii TF-IDF vektorite kui ka SBERT lausevektorite loomisel järgnevates etappides, tagades, et mõlemad sarnasusmeetodid lähtuvad samast informatsiooniallikast

2.4 Kaalutud hinnang mängudele

Arvustuste alusel koostati iga mängu jaoks ühine skoor, mis arvestab korraga nii hinnangute arvu kui ka nende positiivsuse määra. Positiivsete arvustuste osakaalu tähistame p ja arvustuste koguarvu n. Nii liiga väheste kui ka rohkete hinnangute mõju tasakaalustamiseks kohandati IMDb (Internet Movie Database) skeemi. IMDb on ülemaailmne veebipõhine andmebaas filmide, telesaadete ja nendega seotud isikute kohta, kus kogutakse miljoneid kasutajate hinnanguid ning arvustusi. Nad kasutavad seda kaalutud Bayesi hinnangulise keskmise valemit, et tagada skooride stabiilsus ja usaldusväärsus, kuna lihtsalt keskmine võib väheste hinnangute puhul eksitavalt kõrge või madal olla. Valem on:

$$\frac{n}{n+m}p + \frac{m}{n+m}C = score \quad (1)$$

kus C on kogu valimi keskmine positiivsuse määr ja m vastab arvustuste arvu 60. protsentiilile. Nii kaldub suurema hinnete arvuga mängu skoor selle enda positiivsuse suunas, samas kui väheste hinnete korral tõmbab valimi keskmine skoori konservatiivsemale tasemele. Lõpuks normaliseeriti kõik skoorid vahemikku 0–1, et neid saaks otse ja ühtselt kasutada erinevates mudelites ning salvestati veergu *normalized rating*.

2.5 Komponentmudelid

Hübriidmudelite teine oluline komponent lisaks mängu üldisele hinnangule on kahe mängu vaheline sarnasus. Selles töös implementeeriti ja võrreldi kahte erinevat sarnasuskomponenti: traditsioonilist sisupõhist sarnasust TF-IDF vektorite abil ja kaasaegsemat semantilist sarnasust SBERT lausevektorite abil.

2.5.1 Traditsiooniline sisupõhine meetod

Traditsioonilise sisupõhise sarnasuse arvutamiseks teisendati iga mängu kohta varem loodud *combined_text* tekstiväli numbriliseks vektoriks, kasutades TF-IDF meetodit. Selleks kasutati scikit-learn teegi TfidfVectorizer klassi. Vektoriseerimisel:

- Eemaldati inglise keele stoppsõnad.
- Kehtestati terminite minimaalseks esinemissageduseks viis dokumenti (et vältida väga haruldasi ja potentsiaalselt mürarikkaid termineid).
- Piirati sõnavara suurust 5000 kõige sagedamini esineva terminiga (et hoida mälukasutust kontrolli all ning keskenduda olulisematele terminitele).

TF-IDF meetod omistab suurema kaalu terminitele, mis esinevad konkreetses mängukirjelduses sageli, kuid on haruldased kogu mängude korpuses, tuues seeläbi esile mängu iseloomustavad märksõnad. Saadud TF-IDF vektorite maatriksist arvutati iga mängupaari vaheline koosinussarnasus, mis mõõdab vektoritevahelist nurka ja ei sõltu otseselt tekstide pikkusest. Sarnasuste kiireks leidmiseks ehitati scikit-learn teegi NearestNeighbors klassi abil lähimate naabrite indeks, mis võimaldas iga mängu jaoks efektiivselt leida N kõige sarnasemat mängu TF-IDF vektorruumis. See TF-IDF põhine sarnasuskomponent on aluseks esimesele hübriidmudelile.

2.5.2 Semantiline sisupõhine meetod (Sentence Transformer)

Teise, kaasaegsema sarnasuskomponendi loomiseks kasutati semantilist lähenemist, mis põhineb lausevektoritel. Iga mängu *combined_text* väli teisendati tihedaks numbriliseks vektoriks (*embedding*), kasutades Hugging Face sentence-transformers teegi eelkoolitatud mudelit all-MiniLM-L6-v2. See mudel genereerib iga sisendteksti kohta 384-mõõtmelise vektori, kus sarnase tähendusega tekstide vektorid asuvad vektorruumis üksteisele lähemal. Erinevalt TF-IDF-ist, mis keskendub sõnade esinemisele, püüab SBERT tabada teksti tähenduslikku sisu, arvestades sõnade konteksti ja semantilisi seoseid, võimaldades tuvastada sarnasusi ka siis, kui täpsed märksõnad ei kattu.

Arvutatud lausevektorid salvestati ChromaDB in-memory vektorandmebaasi. ChromaDB võimaldab efektiivset sarnasusotsingut (lähimate naabrite leidmist) suurtes vektorhulkades. Kui oli vaja leida konkreetsele mängule semantiliselt sarnaseid mänge, siis esmalt genereeriti selle sisendmängu lausevektor ning seejärel tehti ChromaDB-le päring lähimate naabervektorite leidmiseks. Saadud vektoritevahelised kaugused teisendati koosinussarnasusteks. See semantiline sarnasuskomponent on aluseks teisele hübriidmudelile.

2.6 Hübriidsoovitussüsteemide arhitektuur

Mõlema loodud hübriidsoovitussüsteemi tuumaks on idee kombineerida kahte liiki informatsiooni. Süsteemides ühendatakse mängu üldine hinnang ja populaarsus, mida esindab *normalized_rating*, mängude omavahelise sarnasusega. Omavahelise sarnasus on arvutatud kas sisupõhise (TF-IDF) või semantilise (SBERT) meetodiga.

Lõplik soovitusjärjestus iga seemnemängu jaoks saadakse, arvutades potentsiaalsetele kandidaatmängudele hübriidskoori järgmise valemi alusel:

$$skoor = \alpha \cdot normalized\_rating_{kandidaat} + \beta \cdot sarnasusskoor_{seeme,\, kandidaat} \quad (2)$$

kus:

- $normalized\_rating_{kandidaat}$ on kandidaatmängu kaalutud ja normaliseeritud hinnang nagu seletatud peatükis 2.4.
- $sarnasusskoor_{seeme,\, kandidaat}$ on seemnemängu ja kandidaatmängu vaheline sarnasus, mis arvutatakse ühel kahest viisist, sõltuvalt hübriidmudeli tüübist.
- α ja β on kaalud, mis määravad reitingukomponendi ja sarnasuskomponendi suhtelise olulisuse.

Selle üldise raamistiku alusel implementeeriti käesolevas töös kaks konkreetset hübriidsoovitussüsteemi, mis erinevad teineteisest peamiselt sarnasusskoori arvutamise meetodi poolest, kuid jagavad sama alusloogikat mängu üldise hinnangu (normalized_rating) kaasamisel.

Esimene loodud mudel, sisupõhise sarnasusega hübriid (TF-IDF hübriid), arvutab seemnemängu ja potentsiaalse kandidaatmängu vahelise sarnasusskoori kasutades TF-IDF vektorite koosinussarnasust, nagu kirjeldatud alapeatükis 2.5.1. Selles mudelis kombineeritakse seega mängu üldine populaarsus ja kasutajate hinnangud (*normalized_rating*) selle tekstilisel sisul (märksõnadel ja termide sagedusel) põhineva sarnasusega.

Teine loodud mudel, semantilise sarnasusega hübriid (SBERT hübriid), kasutab sarnasusskoori arvutamiseks all-MiniLM-L6-v2 lausevektorite koosinussarnasust, mida kirjeldati alapeatükis 2.5.2. Siin ühendatakse mängu üldine populaarsus ja hinnangud sarnasusega, mis tugineb teksti sügavamal tähenduslikul sisul, mitte pelgalt sõnade kattuvusel.

Mõlema mudeli puhul genereeritakse seemnemängule soovitused, järjestades kõik teised mängud andmestikus nende arvutatud hübriidskoori alusel kahanevalt ja valides etteantud arv parimaid. Need kaks erineva sarnasuskomponendiga, kuid sama hübriidvalemit ja reitingukomponenti kasutavat mudelit ongi käesoleva töö peamised võrdlusobjektid. Eesmärk on välja selgitada, kas ja millisel määral semantiline sarnasuskomponent suudab pakkuda paremaid või teistsuguseid soovitusi võrreldes traditsioonilisema TF-IDF põhise sisulise sarnasusega, kui mõlemad on integreeritud sarnasesse hübriidraamistikku.

3. Soovitussüsteemide võrdlus ja tulemuste analüüs

Selles alapeatükis kirjeldatakse täpsemalt samme, mis astuti soovitussüsteemide objektiivseks ja korratavaks hindamiseks: testkorpuse moodustamine, tõese sarnasuse defineerimine ning hindamismõõdikute valik.

3.1 Testkorpuse koostamine

Soovituste kvaliteedi hindamiseks moodustati esmalt testkorpus, mis koosnes seemnemängudest (*seed games*). Selleks valiti peatükis 2 kirjeldatud 5000 mängu hulgast 498 mängu kasutades kihistatud valimi meetodit (*stratified sampling*). Valim kihistati kahel alusel:

- Populaarsuse kvantiilid: Mängud jaotati viide kvantiili nende populaarsusskoori alusel
- Peamised žanrid: Igale mängule määrati üks peamine žanr enim levinud žanrite hulgast.

Igast tekkinud kihist (populaarsusgrupp × peamine žanr) valiti proportsionaalne arv mänge, kasutades juhuslikku valikut fikseeritud algväärtusega (*random_state* oli 42), et tagada tulemuste korratavus. See tagas, et igast esindatud kihist valiti vähemalt üks mäng. Selline lähenemine aitas kindlustada, et testkorpus on esinduslik kogu 5000 mängu hõlmava tööandmestiku suhtes ning et testitulemused ei oleks kallutatud vaid teatud tüüpi (väga populaarsete või ühe žanri) mängude suunas. See võimaldas paremini hinnata mudelite üldist toimimist erinevates olukordades.

3.2 Tõese sarnasuse loomine

Kuna käesolevas töös kasutatud andmestikus puudusid individuaalsete kasutajate interaktsioonide andmed, tuli tõese sarnasuse hulk ehk GT (*ground truth*) luua mängude endi tunnuste põhjal. GT defineerib iga seemnemängu jaoks hulga teisi mänge, mida loetakse sellele seemnemängule sarnaseks ja mida soovitussüsteem peaks ideaalis soovitama.

GT põhineb mängude mitmekülgsete tunnuste kogumil, mis hõlmab:

- Žanreid
- Kategooriaid
- Kasutajate lisatud märksõnu

Iga mängu kohta moodustati nende tunnuste alusel unikaalsete, väiketähtedeks normaliseeritud terminite hulk. Kahe mängu vahelist sarnasust mõõdeti seejärel nende tunnuste hulkade Jaccardi indeksi abil:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

kus A ja B tähistavad kahe võrreldava mängu tunnuste hulki. Jaccardi indeks väljendab hulkade kattuvuse määra vahemikus 0 (kattuvus puudub) kuni 1 (hulgad on identsed). Mäng loeti seemnemängule sarnaseks ja lisati selle GT hulka juhul, kui nende tunnuste hulkade Jaccardi indeks ületas eelnevalt kindlaks määratud lävendi. Selline laiapõhjalisem tunnuste kasutamine GT defineerimisel proovib võimaldada tasakaalukamat hinnangut. Kuna sisupõhine mudel on võimekas sarnaste mängude leidmisel märksõnade alusel ning semantiline mudel on potentsiaalikam tuvastama tähenduslikku lähedust, mis võib avalduda suuremas kontekstis, proovisime kasutada GT defineerimisel laiapõhjalisemat lähenemist.

3.3 Hindamismõõdikud

Soovitussüsteemide võrdluses kasutati soovituste nimekirja pikkuse K hindamiseks laialdaselt tunnustatud informatsiooniotsingu mõõdikuid: täpsust (*precision*), saagist (*recall*) ja normaliseeritud kumulatiivset küllastust (*Normalized Discounted Cumulative Gain* ehk nDCG) [11][12]. Need mõõdikud hindavad soovituste kvaliteedi erinevaid aspekte:

- Täpsus: Näitab, kui suur osa soovitatud mängudest olid relevantsed ehk kuulusid seemnemängu GT hulka. Kõrgem väärtus tähendab, et vähem ebaolulisi soovitusi esitati.
- Saagis: Näitab, kui suure osa kõigist GT-s olevatest relevantsetest mängudest suutis mudel etteantud arvu soovitusega üles leida. Kõrgem väärtus tähendab, et vähem relevantseid mänge jäi leidmata.
- nDCG: Hindab soovituste kvaliteeti, arvestades ka relevantsete mängude positsiooni soovituste nimekirjas. Relevantsed mängud, mis asuvad nimekirja alguses, annavad suurema panuse skoori. Tulemus normaliseeritakse ideaalse järjestuse suhtes, andes maksimumväärtuseks 1 täiusliku järjestuse korral.

Nende mõõdikute keskmised väärtused arvutati üle kõigi testkorpuses olevate seemnemängude, et saada üldistatud hinnang mudelite jõudlusele.

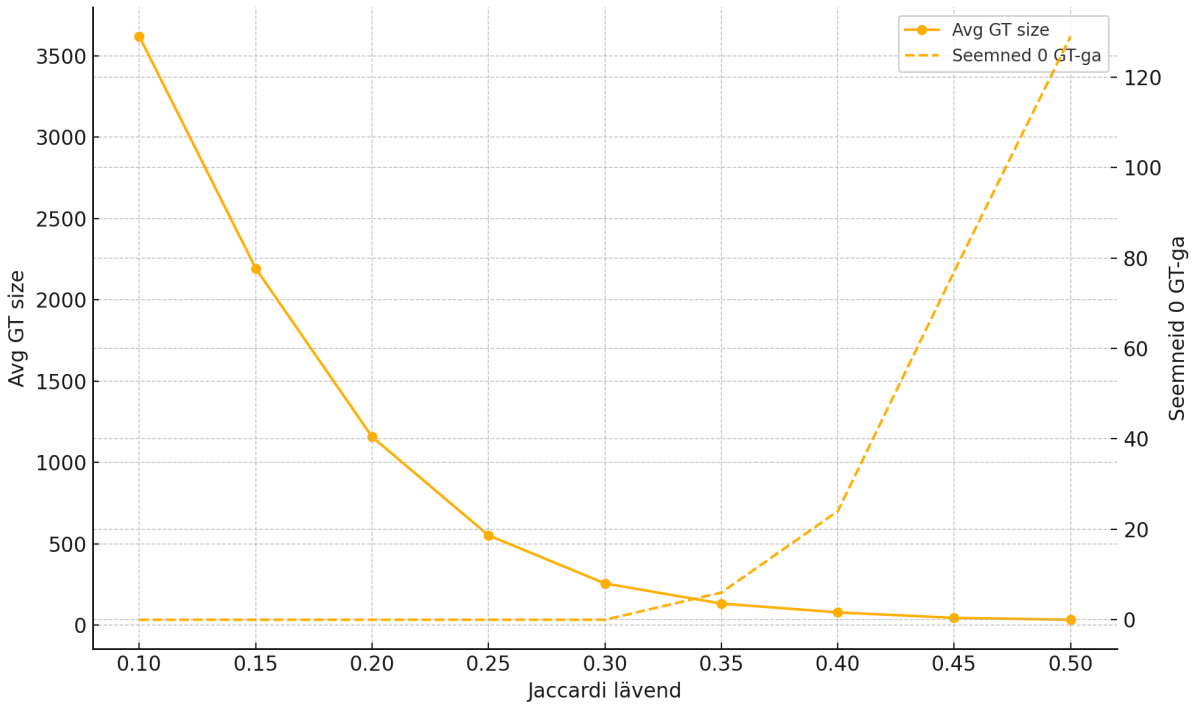
3.4 Jaccardi indeksi lävendi valik

Nagu kirjeldatud alapeatükis 3.2, põhineb tõese sarnasuse määratlemine mängude tunnuste kattuvusel, mida mõõdetakse Jaccardi indeksiga. Sobiva Jaccardi indeksi lävendi valik on GT kvaliteedi seisukohast määrava tähtsusega, kuna see mõjutab otseselt GT komplektide suurust, sisu ja seeläbi ka mudelite hindamisel saadavaid mõõdikute väärtusi. Liiga madal lävend võib GT-sse lisada palju ebaolulisi vasteid, samas kui liiga kõrge lävend võib GT jätta liiga väikeseks või isegi tühjaks paljude seemnemängude puhul.

Optimaalse lävendi leidmiseks viidi läbi eksperimentaalne analüüs. GT genereeriti ja mõlema hübriidmudeli jõudlust (täpsus, saagis ja nDCG) hinnati üheksa erineva Jaccardi lävendi väärtuse juures: 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45 ja 0.50. Iga testitud lävendi jaoks oli soovituste nimekirja ehk K pikkuseks 10 ning registreeriti järgmised näitajad:

- Keskmine GT komplekti suurus (*avg_gt_size*) üle kõigi seemnemängude.
- Nende seemnemängude arv, mille GT komplekt jäi tühjaks (*seeds_w_zero_gt*).
- Hindamismõõdikud täpsus, saagis ja nDCG mõlema, nii sisupõhise kui ka semantilise, hübriidmudeli jaoks.

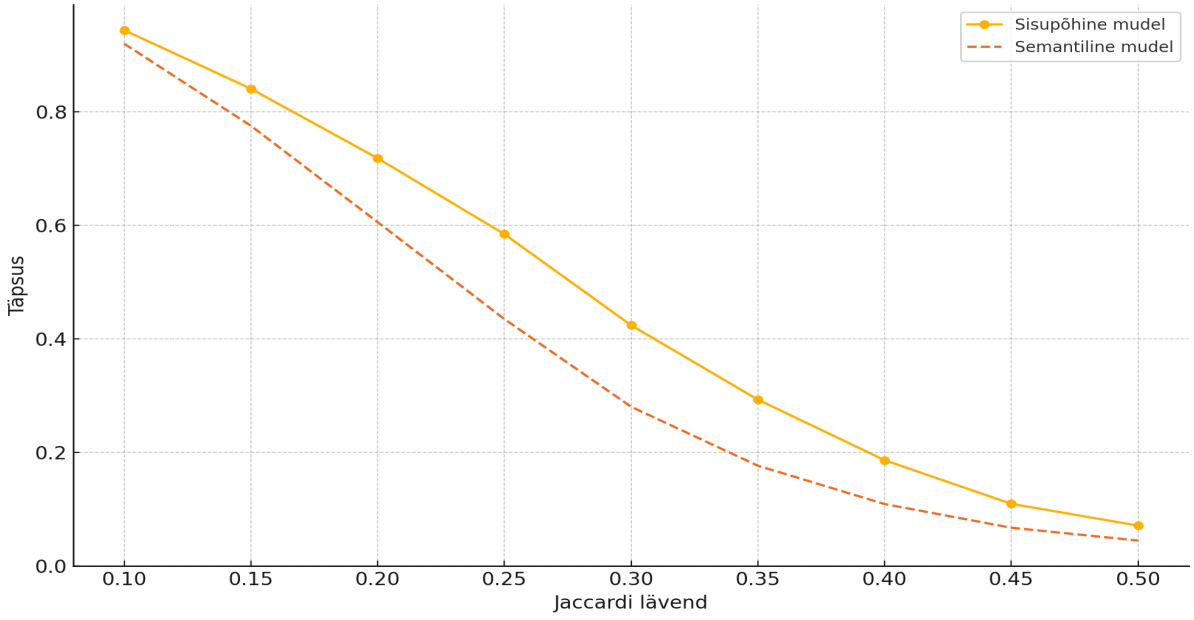
Selle analüüsi tulemused on koondatud Tabelisse 1 ning illustreeritud Joonistel 1 kuni 4.



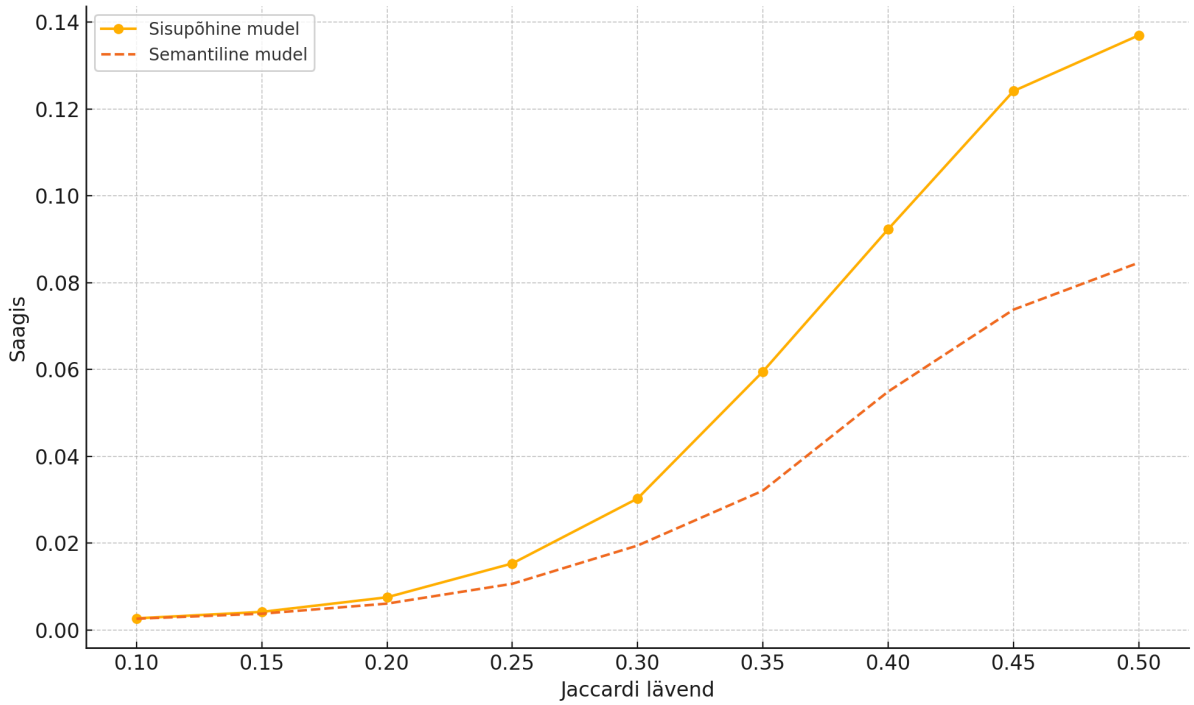
Joonis 1. Keskmise GT suuruse ja 0-GT-ga seemnete arvu sõltuvus Jaccardi indeksi lävendist

Tabel 1. Jaccardi indeksi lävendi mõju GT karakteristikutele ja mudelite jõudlusele (K=10)

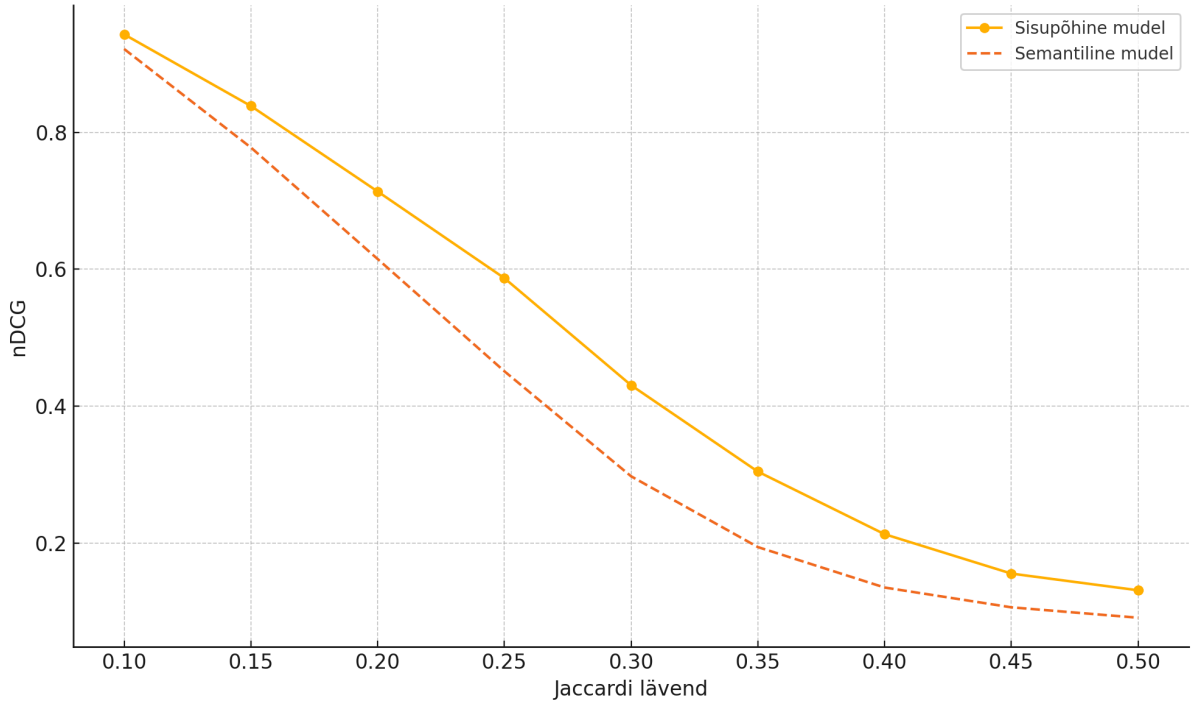
Läv end	Keskm. GT suurus	Seemneid 0 GT-ga	Täpsus (Sisupõhine)	Saagis (Sisupõhine)	nDCG (Sisupõhine)	Täpsus (Semantiline)	Saagis (Semantiline)	nDCG (Semantiline)
0.1	3618.67	0	0.944	0.003	0.943	0.92	0.003	0.922
0.15	2192.66	0	0.841	0.004	0.838	0.775	0.004	0.778
0.2	1156.66	0	0.718	0.008	0.714	0.606	0.006	0.615
0.25	551.57	0	0.584	0.015	0.587	0.435	0.011	0.451
0.3	256	0	0.424	0.03	0.43	0.28	0.019	0.297
0.35	132.37	6	0.293	0.06	0.304	0.177	0.032	0.194
0.4	78.01	24	0.186	0.092	0.213	0.109	0.055	0.135
0.45	44.34	77	0.11	0.124	0.155	0.067	0.074	0.106
0.5	32.94	129	0.071	0.137	0.131	0.045	0.085	0.091



Joonis 2. Mõlema hübriidmudeli Täpsus sõltuvalt Jaccardi indeksi lävendist.



Joonis 3. Mõlema hübriidmudeli Saagis sõltuvalt Jaccardi indeksi lävendist.



Joonis 4. Mõlema hübriidmudeli nDCG sõltuvalt Jaccardi indeksi lävendist.

Esitatud andmetest (Tabel 1) ja joonistelt ilmneb selge seos Jaccardi indeksi lävendi väärtuse ning nii GT karakteristikute kui ka mudelite mõõdikute vahel. Lävendi langetamisel alla 0.20 kasvab keskmine GT suurus märkimisväärselt, ulatudes tuhandetesse. See viitab sarnasuse definitsiooni liigsele leebusele, mille tulemusena GT ei ole piisavalt eristusvõimeline. Kuigi täpsusnäitajad on selliste madalate lävendite korral kõrged, on saagisenäitajad äärmiselt madalad. See on ka ootuspärane, kuna soovituste nimekirja suurus oli ainult 10 ning seetõttu on raske GT hulgast olulist osa katta.

Lävendi tõstmisel hakkab GT keskmine suurus vähenema ja saagis suurenema. Samas, alates lävendist 0.35, hakkab suurenema ka nende seemnemängude arv, millel puuduvad GT hulgas vasted. Eesmärgiks oli leida tasakaalupunkt, kus GT on piisavalt fokusseeritud, et võimaldada mudelite adekvaatset hindamist, kuid samas piisavalt ulatuslik, et minimaliseerida tühjade GT hulkadega seemnemängude arvu. Seega valiti edasisteks katseteks ja mudelite lõplikuks hindamiseks Jaccardi indeksi lävendiks 0.35

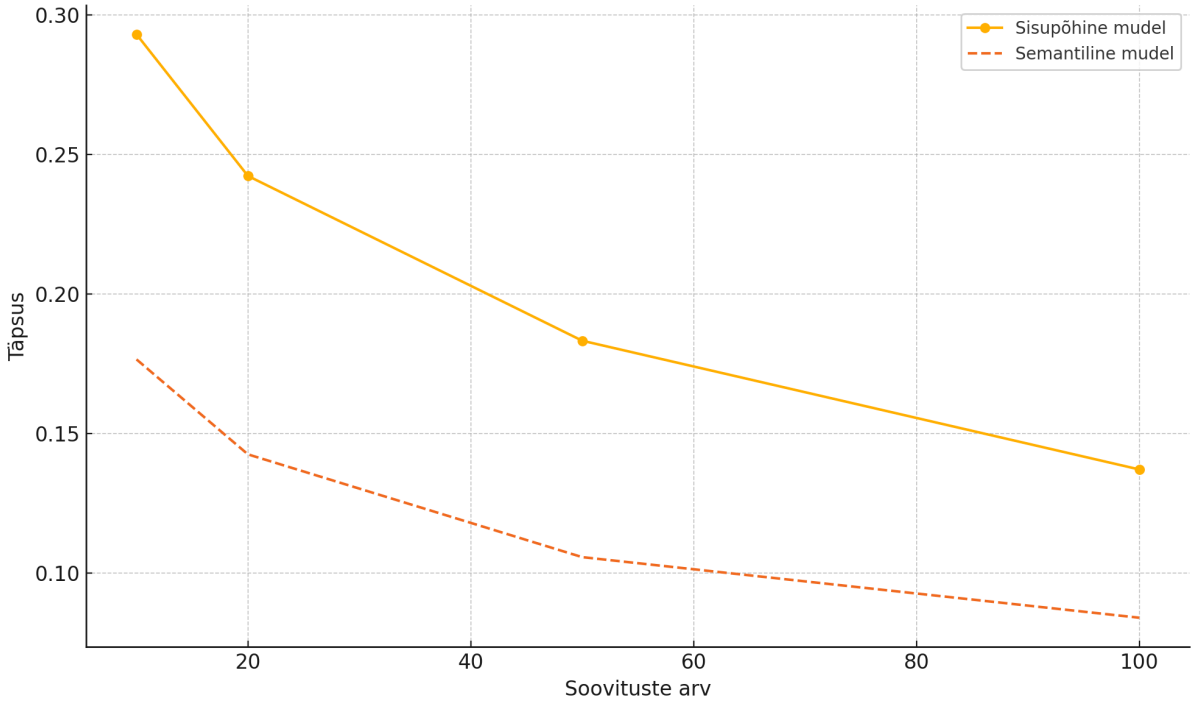
3.5 Hübriidmudelite kvantitatiivne võrdlus

Pärast sobiva Jaccardi indeksi lävendi (0.35) leidmist ja selle alusel lõpliku tõese sarnasuse hulga genereerimist, hinnati mõlema hübriidsoovitussüsteemi jõudlust. Hindamine viidi läbi erinevate soovituste arvu K väärtuste juures: 10, 20, 50 ja 100. Mõõdikutena kasutati, nagu eelnevalt peatükis 3.3 kirjeldatud, täpsust, saagist ja normaliseeritud kumulatiivset küllastust. Hübriidmudelite parameetrid α ja β hoiti konstantsena (α=0.6,β=0.4) kogu testimise vältel. Eesmärgiks oli leida tasakaal, mis arvestaks nii kasutajate üldist tagasisidet (läbi *normalized_rating* välja) kui ka sisulist või semantilist lähedust ning need α ja β väärtused pakkusid seda.

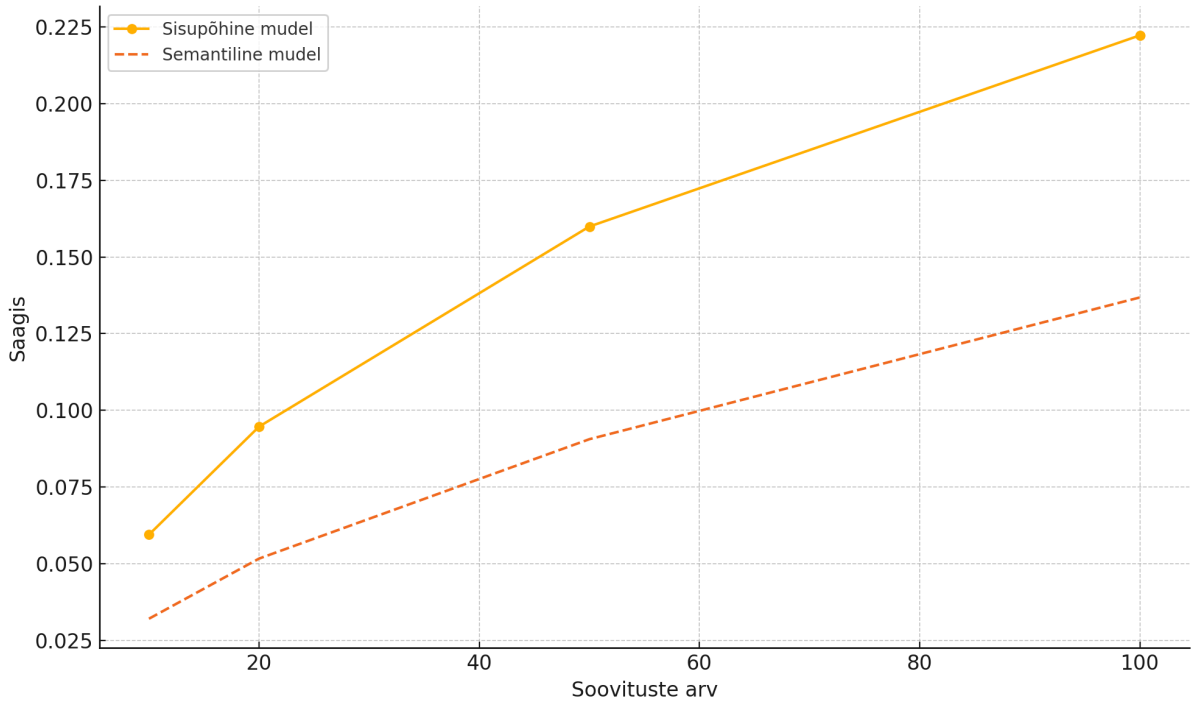
Saadud tulemused on koondatud Tabelisse 2 ning visualiseeritud Joonistel 5 kuni 7.

Tabel 2. Hübriidmudelite jõudlusnäitajad erinevate K väärtuste juures

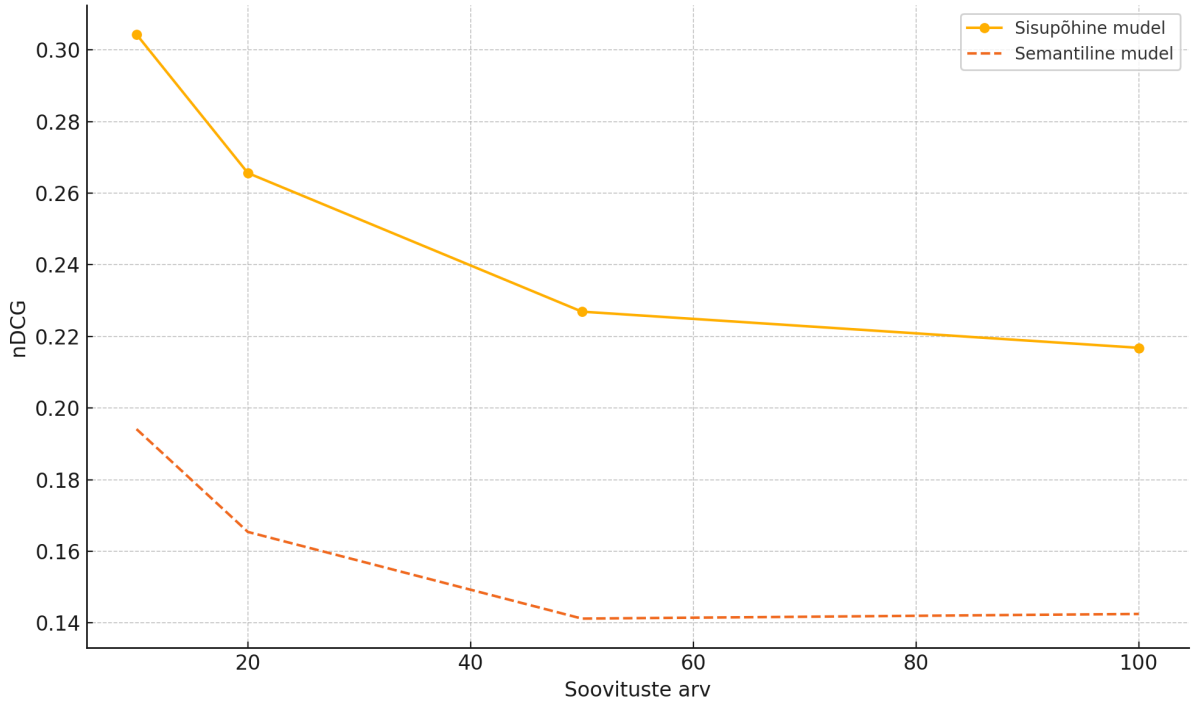
K	Mudel	Täpsus	Saagis	nDCG
10	Sisupõhine	0.293	0.0595	0.3043
10	Semantiline	0.1766	0.0321	0.1941
20	Sisupõhine	0.2423	0.0947	0.2656
20	Semantiline	0.1426	0.0517	0.1654
50	Sisupõhine	0.1833	0.1599	0.2269
50	Semantiline	0.1058	0.0906	0.1412
100	Sisupõhine	0.1372	0.2222	0.2168
100	Semantiline	0.0841	0.1368	0.1425



Joonis 5. Mudelite täpsus sõltuvalt K väärtusest



Joonis 6. Mudelite saagis sõltuvalt K väärtusest



Joonis 7. Mudelite nDCG sõltuvalt K väärtusest

3.6 Tulemuste analüüs ja järeldused

Tabelist 2 ja joonistelt 5-7 ilmneb selge ja järjepidev trend: sisupõhine hübriidmudel edestab semantilist hübriidmudelit kõigi testitud K väärtuste juures ning seda kõigi kolme mõõdiku osas. See tähelepanek on mõnevõrra ootamatu, arvestades semantiliste mudelite potentsiaali tabada sügavamaid tähenduslikke seoseid tekstides.

Vaadeldes mõõdikute käitumist K väärtuse muutudes:

- Täpsus (Joonis 5): Mõlema mudeli puhul langeb täpsus ootuspäraselt K väärtuse kasvades. See on loogiline, kuna pikemates soovituste nimekirjades on raskem hoida kõrget asjakohaste vastete osakaalu. Sisupõhise mudeli täpsus on aga igal K väärtusel märgatavalt kõrgem kui semantilise mudeli oma.
- Saagis (Joonis 6): Mõlema mudeli saagis kasvab K väärtuse suurenedes, mis on samuti ootuspärane. Pikemad soovituste nimekirjad võimaldavad katta suurema osa GT-s olevatest relevantsetest mängudest. Ka siin on sisupõhine mudel semantilisest järjepidevalt parem. Kui K=100, saavutab sisupõhine mudel saagise 0.222, semantiline aga 0.137. See tähendab, et isegi 100 soovituse puhul leiab sisupõhine mudel GT-st üles umbes 22% relevantsetest mängudest, semantiline aga umbes 14%.
- nDCG (Joonis 7): See mõõdik, mis arvestab ka asjakohaste vastete järjestust, peegeldab täpsuse trendi. Sisupõhise mudeli nDCG on kõrgem, mis viitab sellele, et see mitte ainult ei leia rohkem relevantseid mänge, vaid paigutab need ka soovituste nimekirjas keskmiselt kõrgematele positsioonidele kui semantiline mudel.

Nagu esitatud tulemustest selgub, edestas sisupõhist TF-IDF komponenti kasutav hübriidmudel järjepidevalt semantilist hübriidmudelit kõigi kolme hindamismõõdikuning kõigi testitud K väärtuste lõikes. See tähelepanek viitab mitmele potentsiaalsele tegurile, mis võisid antud eksperimendis sisupõhise lähenemise paremust soodustada.

Kuigi tõese sarnasuse loomisel kasutati mitmekülgseid tunnuseid ja Jaccardi indeksit, põhines GT siiski selgesõnaliste tunnuste kattuvusel. TF-IDF meetod on oma olemuselt optimeeritud just selliste märksõnade ja nende esinemissageduste tabamiseks. Kuna *combined_text* väli, mida TF-IDF analüüsis, sisaldas neid samu tunnuseid ka tekstilisel kujul, oli sellel meetodil eelis GT-s defineeritud sarnasuse äratundmisel. Semantiline mudel, ei pruukinud selle GT suhtes oma tugevusi täiel määral realiseerida, kui GT hindas peamiselt just tunnuste otsest jagamist.

Samuti võisid Steam'i mängude puhul olla just konkreetsed tunnused nagu žanrid, kategooriad ja kasutajate poolt antud märksõnad sageli peamised ja väga informatiivsed. On võimalik, et mängude pikemad tekstilised kirjeldused ei lisanud piisavalt eristavat semantilist informatsiooni ning seetõttu ei suutnud anda mida TF-IDF poolt kasutatavad selged tunnused juba ei katnud.

Lisaks ei pruukinud valitud üldotstarbeline semantiline mudel (all-MiniLM-L6-v2) olla piisavalt spetsialiseerunud videomängude alal, et ületada TF-IDF robustsust selgelt defineeritud tunnuste kontekstis. TF-IDF lihtsus ja keskendumine termide statistilisele olulisusele võisid antud juhul osutuda eeliseks.

Seega, kuigi semantilised mudelid pakuvad teoreetiliselt sügavamat tekstimõistmist, näitas käesolev uurimus, et konkreetse andmestiku ja ülesandepüstituse juures võib traditsioonilisem, märksõnade kattuvusel põhinev sisupõhine meetod anda paremaid tulemusi. See rõhutab vajadust hoolikalt kaaluda kõiki neid aspekte soovitussüsteemi arendamisel ja hindamisel.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli võrrelda kahte hübriidsoovitussüsteemi Steam videomängude andmestikul: traditsioonilisel TF-IDF sisupõhisel sarnasusel põhinevat mudelit ning kaasaegsemat, SBERT lausevektoreid kasutavat semantilist mudelit. Mõlemad sarnasuskomponendid kombineeriti mängude üldise kaalutud hinnanguga, kasutades fikseeritud kaalusid (α=0.6 reitingule, β=0.4 sarnasusele). Uurimuse aluseks oli 5000 populaarseimast ingliskeelsest Steam mängust koosnev valim, mis pärines märts 2025 seisuga uuendatud andmestikust.

Mudelite hindamiseks loodi 498 seemnemängust koosnev testkorpus ning tõese sarnasuse (GT) hulk defineeriti mängude mitmekülgsete tunnuste (žanrid, kategooriad, kasutajate tag'id) kattuvuse alusel, kasutades Jaccardi indeksit (lävendiga 0.35). Süsteemide jõudlust mõõdeti täpsuse (*precision*), saagise (*recall*) ja nDCG näitajatega, kus K=[10,20,50,100].

Eksperimendi tulemused näitasid järjepidevalt, et sisupõhine (TF-IDF) hübriidmudel edestas semantilist (SBERT) hübriidmudelit kõigi mõõdikute ja testitud K väärtuste lõikes. Järeldati, et sisupõhise mudeli paremus tulenes tõenäoliselt GT olemusest, mis premeeris selgesõnaliste tunnuste kattuvust, Steam mängude andmestiku spetsiifikast (kus konkreetsed tag'id on olulised) ning kasutatud üldotstarbelise semantilise mudeli piirangutest antud domeenis. Töö rõhutab, et soovitussüsteemide efektiivsus sõltub tugevalt kontekstist ning metoodikavalikutest, ja et traditsioonilised meetodid võivad teatud tingimustel anda paremaid tulemusi kui kaasaegsemad alternatiivid. Edasine uurimistöö võiks keskenduda domeenispetsiifilisematele semantilistele mudelitele ja alternatiivsetele GT loomise viisidele.

Viited

[1] R. Yin, K. Li, G. Zhang, and J. Lu, "A deeper graph neural network for recommender systems," *Knowledge-Based Systems*, vol. 185, p. 105020, Dec. 2019, https://doi.org/10.1016/j.knosys.2019.105020

[2] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data*, vol. 9, no. 1, pp. 1–36, May 2022, https://doi.org/10.1186/s40537-022-00592-5

[3] J. Gong, Y. Ye, and K. Stefanidis, "A Hybrid Recommender System for Steam Games," in *Communications in Computer and Information Science*, Cham: Springer International Publishing, 2020, pp. 133–144. doi: https://doi.org/10.1007/978-3-030-44900-1_9.

[4] M. Srifi, A. Oussous, A. Ait Lahcen, and S. Mouline, "Recommender Systems Based on Collaborative Filtering Using Review Texts—A Survey," *Information*, vol. 11, no. 6, p. 317, Jun. 2020, https://doi.org/10.3390/info11060317

[5] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep Learning Based Recommender System: A Survey and New Perspectives," *ACM Computing Surveys*, vol. 52, no. 1, pp. 1–38, Feb. 2019, https://doi.org/10.1145/3285029

[6] K. Benabbes, K. Housni, A. E. Mezouary, and A. Zellou, "Recommendation System Issues, Approaches and Challenges Based on User Reviews," *Journal of Web Engineering*, Apr. 2022, https://doi.org/10.13052/jwe1540-9589.2143

[7] Z. Zhao *et al.*, "Recommender Systems in the Era of Large Language Models (LLMs)," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 6889–6907, Nov. 2024, https://doi.org/10.1109/tkde.2024.3392335

[8] L. Zheng, V. Noroozi, and P. S. Yu, "Joint Deep Modeling of Users and Items Using Reviews for Recommendation," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, Feb. 2017, pp. 425–434. https://doi.org/10.1145/3018661.3018665

[9] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural Attentional Rating Regression with Review-level Explanations," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, New York, New York, USA: ACM Press, 2018, pp. 1583–1592, https://doi.org/10.1145/3178876.3186070

[10] J. Chicaiza and P. Valdiviezo-Diaz, “A Comprehensive Survey of Knowledge Graph-Based Recommender Systems: Technologies, Development, and Contributions,” *Information*, vol. 12, no. 6, p. 232, May 2021, https://doi.org/10.3390/info12060232

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[12] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of IR techniques,” *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, Oct. 2002, https://doi.org/10.1145/582415.582418

[13] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian, and X. Xie, “Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations,” *ACM Transactions on Information Systems*, Apr. 2025, https://doi.org/10.1145/3731446

[14] J. Chen et al., “When large language models meet personalization: perspectives of challenges and opportunities,” World Wide Web, vol. 27, no. 4, Jun. 2024, https://doi.org/10.1007/s11280-024-01276-1

[15] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, Jan. 1988, https://doi.org/10.1016/0306-4573(88)90021-0

[16] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, doi: https://doi.org/10.18653/v1/d19-1410

Lisa 1. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Mihkel Järviste**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose,
   "**Sisupõhiste ja semantiliste vektorkujutustest hübriidmudelite võrdlus e-kaubanduse soovitussüsteemides**"
   mille juhendajad on Andres Järviste ja Margus Niitsoo, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi 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.

Mihkel Järviste

15.05.2025

Lisa 2. Lähtekood

Praktilise osa lähtekood (GitHub):

https://github.com/mjarviste/ai_semantic_recommender