EESTI
STATISTIKA

# Kaupade klassifitseerimine tekstilise kirjelduse alusel: rakendusuuring

Majandus- ja kommunikatsiooniministeeriumi ja Statistikaameti koostöökokkulepe nr 1.9-8/22-621-1

Hans Hõrak

28.06.2024

Statistikaamet

2024

# Sisukord

# Sissejuhatus

Kombineeritud nomenklatuur on Euroopa Liidus kasutatav kaupade klassifikaator. Iga imporditav ja eksporditav objekt või koorem peaks kuuluma ühe konkreetse 8-kohalise koodi alla. Koodisüsteem on olemuselt hierarhiline (tabel 1), kusjuures hierarhia kolm esimest taset (6 esimest numbrit) ühilduvad globaalse harmoniseeritud koodisüsteemiga (HS-kood).

**Tabel 1. Rukkileiva klassifitseerimine nomenklatuuri hierarhias ning klassifikaatori kategooriate arvud eri tasemetel**

| Tase | Kood | Kirjeldus nomenklatuuris | # kat 1* | # kat 2** |
|---|---|---|---|---|
| **HS2** | 19 | Tooted teraviljast, jahust, tärklisest või piimast; valikpagaritooted | 98 | 97 |
| **HS4** | 1905 | Leiva- ja saiatooted, valikpagaritooted, koogid, küpsised jms pagaritooted, kakaoga või kakaota; armulaualeib, tühjad kapslid farmaatsiatööstusele, oblaadid, riispaber jms | 1211 | 1000 |
| **HS6** | 190590 | Leiva- ja saiatooted, valikpagaritooted, koogid, küpsised jms pagaritooted, kakaoga või kakaota; armulaualeib, tühjad kapslid farmaatsiatööstusele, oblaadid, riispaber jms (v.a kuivikleib, piparkoogid jms, magusad küpsised, vahvlid veesisaldusega kuni 10% massist, kuivikud jms röstitud leiva- ja saiatooted) | 5255 | 3204 |
| **CN8** | 19059030 | Leiva- ja saiatooted, mis ei sisalda mett, mune, juustu, marju või puuvilju ning mis sisaldavad või ei nii suhkrut kui rasva kuni 5% kuivaine massist | 8789 | 4734 |

***** Kategooriate arv **käsitletava** koodi tasemel Eesti Intra- ja Extrastat kombineeritud andmestikus aastatel 2019–2023, kus on vähemalt 1 unikaalne kirjeldus minimaalse pikkusega 6 tähemärki.

****** Kategooriate arv **käsitletava** koodi tasemel Eesti Intra- ja Extrastat kombineeritud andmestikus perioodil 2019-2023, kus on vähemalt 30 unikaalset kirjeldust minimaalse pikkusega 6 tähemärki (masinõppe eeldustele vastav osa nomenklatuurist)

Olukordades, kus on vaja määrata kaubakood, võib olla kättesaadav kaupa kirjeldav tekst (näiteks kaubaga kaasnevast dokumentatsioonist, kirjadest pakendil või kaubale peale vaadates seda iseloomustav kirjeldus). Käesoleva uuringuga püütakse välja selgitada, kas ja millise täpsusega on võimalik välja arendada masinõppe süsteem, mis suudaks kirjelduse tekstist tuletada kombineeritud nomenklatuuri koodi või vähemalt otsingut kitsendada. Lisaks hinnatakse sellise süsteemi välja arendamise mõistlikkust lähtuvalt võimalikust saavutatavast kasutegurist: kui suurele osale andmestikust suudaks süsteem iseseisvalt või operaatori tööriistana määrata õige kaubakoodi?

## Kirjanduse ülevaade

Kõik riigid, mis tegelevad väliskaubandusega, tunnevad ilmselt vajadust kaupade automaatse klassifitseerimise järele, kuna kaupu on palju, eri kaupadel on erinevad maksumäärad ning riigid huvituvad kaubavahetuse struktuurist (mida kui palju kuhu eksporditakse ja mida kui palju kuskilt imporditakse). Seetõttu on praeguseks avaldatud mitmeid teadusartikleid, kus on püütud lahendada sama või sarnast probleemi: kaubakirjelduste automaatset klassifitseerimist kaubakoodide alla (tabel 2). Kõik leitud uuringud peale ühe keskenduvad üksikutele nomenklatuuri HS2 kategooriatele, püüdmata automatiseerida kaubakoodi klassifitseerimist kogu nomenklatuuri ulatuses.

**Tabel 2. Kirjanduse ülevaate metaanalüüsi tabel**

| Viide | Kogum | Algoritm | Andmestiku suurus | Klasside arv | Täpsus |
|---|---|---|---|---|---|
| **Jahanshahi** *et al*., **2021** | toiduained | Bi-LSTM | 4026 | 44 | 91,2 |
| **He** *et al*., **2021** | 84 | Bert+CNN | 152 183 | 105 | 87,4 |
| **Ding** *et al*., **2015** | 90 (optika) | Background net | 83 830 | 204 | 74,9 |
| **Du** *et al*., **2021** | Tekstiilitööstu-sega seonduv | HSCodeNet | 165 416 | 899 | 77,2 |
| **Lee** *et al*., **2024** | 84, 85, 90 | KLUE-RoBERTA | 206 435 | 925 | 86,1 |
| **Binh** *et al*., **2021** | 03, 62, 85, 30, 38, 40 | Att-RNN | 751 329 | 1167 | 67,1 |
| **Paramartha** *et al*., **2021** | min 100 kirjega kaubad | Multinomial-NB | 276 917 | 1864 | 71,7 |
| **Chen** *et al*., **2021** | Kõik | HLR, LSTM, NMT-HL | 476 128 | 4257 | 46 |

Võrreldes uuringuid klassifitseeritavate kaubakoodide arvu ning saavutatava täpsuse poolest, ilmneb selge seos: mida suuremat osa klassifikaatorist püüda automatiseerida, seda madalam täpsus saavutatakse (klasside arvu ja täpsuse korrelatsioon r = -0,9, p = 0,002). Uuringud toovad olulise probleemina välja klasside ebaühtlase esinemissageduse: kui mingit tüüpi kaup satub tolli vaid kaks korda aastas ning kirjeldus on olemas vaid ühel neist, siis selle kauba kohta ei saa olla piisavalt treeningandmeid, et masinõppemudel suudaks seda kaupa ära tunda sama hästi kui teist kaupa, mille kohta tekib iga päev mitu näidet. Ding ja kolleegid (2015) rõhutavad ka suure klasside arvu ja vähese eristava info probleemi: kirjeldused on tihti nii napisõnalised, et neid ei olegi võimalik ühe konkreetse koodiga seostada. Näiteks sõna „liha“ võib viidata kana-, veise-, lamba- või sealiha hakitud ja hakkimata, värsketele ja sügavkülmutatud variantidele. Ilmselt sellel põhjusel saavutavad ka suhteliselt lihtsad masinõppe algoritmid sarnaseid tulemusi sügavate tehisnärvivõrkude ja keeletehnoloogiatega – andmete ja masinõppe ülesande iseloom on selline, et „võimsamate“ meetodite tüüpilistest eelistest ei saa kasu lõigata. See ilmneb Cheni ja kolleegide (2021) töös, kus hierarhiline multinomiaalne logistiline regressioon saavutab pisut suurema täpsuse

kui masintõlke keelemudel ja LSTM-tehisnärvivõrk. Selle klassifitseerimisülesande puhul ei olegi ehk vaja mudelit, mis mõistaks keelt ja tähendust, vaid sellist, mis suudaks väga spetsiifiliste sõnade ja sõnakombinatsioonide põhjal kirjeid eristada. Suuremad keelemudelid on ehk võimelised õppima selgeks *karburaatori* tähenduse ja konteksti mitmes keeles, aga see ei ole kaubakoodi klassifitseerimiseks ilmtingimata vajalik. Lee jt (2024) toovad probleemina välja ka „muu“ või „varia“ kategooriad, mida esineb paljude HS6 tasemete puhul – sellised kaubad, mis ei lähe teiste sama rubriigi kaupadega päris ühe koodi alla, aga mille jaoks ei saa või ei ole mõtet eraldi CN8 kategooriat teha. Need kategooriad on kohati üpris suure esindusega (võibolla eelistataksegi kaubale määrata pigem varia kategooria, kui riskida konkreetsema, aga vale koodi panemisega), aga suurema heterogeensuse tõttu on nende kategooriate klassifitseerimise täpsus konkreetsemate kategooriatega võrreldes väiksem. Kokkuvõttes on viimastel aastatel tekkinud suur huvi kaubakoodide automaatse klassifitseerimise vastu ning on proovitud mitmeid väga kompleksseid algoritme, aga kogu klassifikaatorit aktsepteeritava täpsusega automatiseerida ei ole suudetud.

## Eesti väliskaubanduse andmete kaubakirjelduste analüüs

Eestis kogutakse väliskaubanduse andmeid kahel viisil: Intrastatis kajastuvad Euroopa Liidu sisesed tehingud (valimis on sõltuvalt lävendist suuremad kauplejad), Extrastatis kajastuvad tollis registreeritud tehingud kolmandate riikidega. Eesti Intrastati andmestikus on viimase viie aasta jooksul tekkinud kokku 40,7 miljonit kirjet 9083 erineva CN8 kaubakoodiga (tabel 3). Kauba kirjelduse väli on Intrastatis täidetud 40%-l kirjetest. Extrastati (tabel 4) andmestiku puhul tuleb kõrvale jätta kaubakood 99510000, mille alla lubatakse lihtsustatud korras registreerida kõiksuguseid e-kaubanduse väiketehinguid. Vaadates adekvaatselt kodeeritud osa andmestikust, on Extrastat oluliselt väiksema andmemahuga, aga kauba kirjelduse väli on täidetud peaaegu kõigil kirjetel.

**Tabel 3. Intrastat: kirjete ja CN8 kaubakoodide arv**

| Aasta | Kogu Intrastati andmestik | | Kaubad, millel on min 6 sümboliga* kirjeldus | | Kaubad, millel on MIN 30 unikaalset** min 6 sümboliga kirjeldust* | |
|---|---|---|---|---|---|---|
| | N kirjeid | N koode | N kirjeid | N koode | N kirjeid | N koode |
| **2019** | 7 422 178 | 7830 | 3 039 577 | 6080 | 373 766 | 3084 |
| **2020** | 7 351 443 | 7757 | 2 842 868 | 6112 | 347 949 | 3081 |
| **2021** | 7 783 396 | 7816 | 2 956 333 | 6185 | 248 477 | 3076 |
| **2022** | 8 708 957 | 7850 | 3 478 184 | 6333 | 776 938 | 3075 |
| **2023** | 9 441 022 | 7796 | 3 963 253 | 6235 | 756 130 | 3030 |
| **Kokku** | **40 706 996** | **9083** | **16 280 215** | **7803** | **2 503 260** | **3264** |
| **% kokku** | **100%** | **100%** | **40%** | **85,9%** | **6,1%** | **35,9%** |

* Pärast puhastamist (kirjavahemärkide ja tundmatute sümbolite eemaldamine ning suurtähtede väikseks muutmine).

** 30 unikaalset kirjeldust kogu viieaastase ajavahemiku kohta, mitte aastate kaupa eraldi. Justkui moodustaks masinõppe andmestikku sellest 5-aastasest ajavahemikust.

**Tabel 4. Extrastat: kirjete ja CN8 kaubakoodide arv**

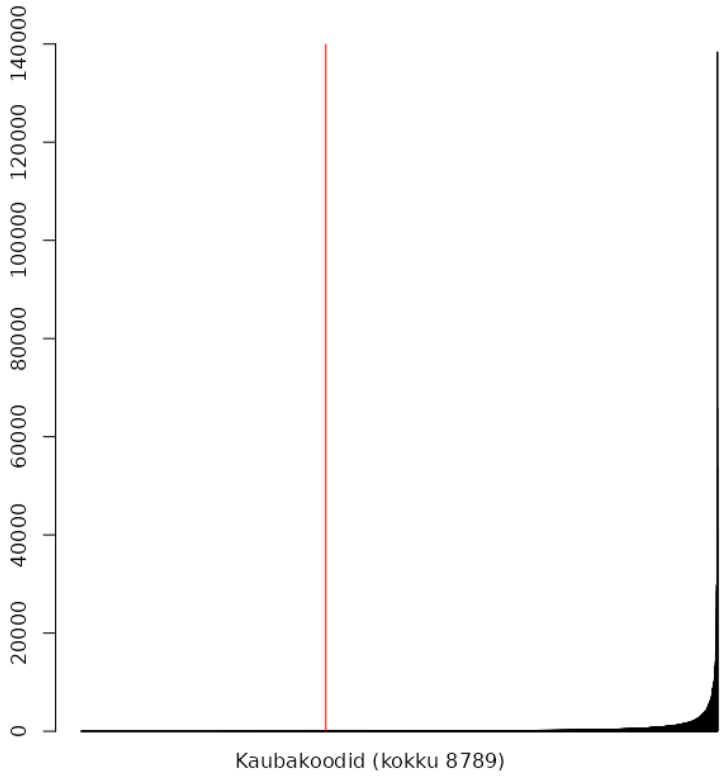
| Aasta | Kogu Extrastati andmestik ilma koodita 99510000 | | Kaubad, millel on min 6 sümboliga kirjeldus | | Kaubad, millel on min 30 unikaalset min 6 sümboliga kirjeldust | |
|---|---|---|---|---|---|---|
| | N kirjeid | N koode | N kirjeid | N koode | N kirjeid | N koode |
| **2019** | 756 773 | 6159 | 750 524 | 6151 | 239 263 | 3543 |
| **2020** | 708 089 | 6151 | 702 092 | 6133 | 206 050 | 3554 |
| **2021** | 948 995 | 6580 | 940 934 | 6560 | 271 194 | 3578 |
| **2022** | 822 500 | 6534 | 815 954 | 6523 | 216 412 | 3568 |
| **2023** | 728 066 | 6446 | 722 270 | 6436 | 231 688 | 3535 |
| **Kokku** | **3 964 423** | **8132** | **3 931 774** | **8115** | **1 164 607** | **3743** |
| **% kokku** | **100%** | **100%** | **99,2%** | **99,8%** | **29,4%** | **46%** |

Ühendades andmestikud (tabel 5), saame ettekujutuse sellest, kui suur osa klassifikaatorist oleks teoreetiliselt automatiseeritav: poolte viie aasta jooksul esinenud kaubakoodide kohta on vähemalt 30 unikaalset vähemalt kuuesümbolilist kirjeldust. 556 kaubakoodi on sellised, mille kohta pole viie aasta jooksul tekkinud ühtegi kaubakirjeldust. Kuna masinõppeks on vaja unikaalseid kirjeldusi piisava esindatusega kaubakoodide juurde, saab lõpuks masinõppe andmestikku kaasata üpris väikese osa kogu andmestikust. Kaupade ekstreemset ebatasakaalu illustreerib joonis 1.

**Tabel 5. Kombineeritud andmestik: kirjete ja CN8 kaubakoodide arv**

| Aasta | Kombineeritud Intra- ja Extrastat ilma Extrastati koodita 99510000 | | Kaubad, millel on min 6 sümboliga kirjeldus | | Kaubad, millel on min 30 unikaalset min 6 sümboliga kirjeldust | |
|---|---|---|---|---|---|---|
| | N kirjeid | N koode | N kirjeid | N koode | N kirjeid | N koode |
| **2019** | 8 178 951 | 8061 | 3 790 101 | 7080 | 600 550 | 4478 |
| **2020** | 8 059 532 | 7997 | 3 544 960 | 7070 | 548 611 | 4473 |
| **2021** | 8 732 391 | 8120 | 3 897 267 | 7323 | 513 399 | 4493 |
| **2022** | 9 531 457 | 8185 | 4 294 138 | 7455 | 987 984 | 4511 |
| **2023** | 10 169 088 | 8129 | 4 685 523 | 7381 | 987 143 | 4484 |
| **Kokku** | **44 671 419** | **9345** | **20 211 989** | **8789** | **3 637 687** | **4734** |
| **% kokku** | **100%** | **100%** | **45,2%** | **94,1%** | **8,1%** | **50,6%** |

**Joonis 1. Ekstreemne ebatasakaal kaupade esindatuses. Unikaalseid minimaalselt kuue märgi pikkuseid kirjeldusi kaubakoodi kohta aastatel 2019–2023 kombineeritud andmestikus. Punane joon tähistab piiri, millest alates on kaubakoodide kohta vähemalt 30 unikaalset kirjeldust (selliseid kaupu on 4734). Jaotuse Gini koefitsient on 0,866**

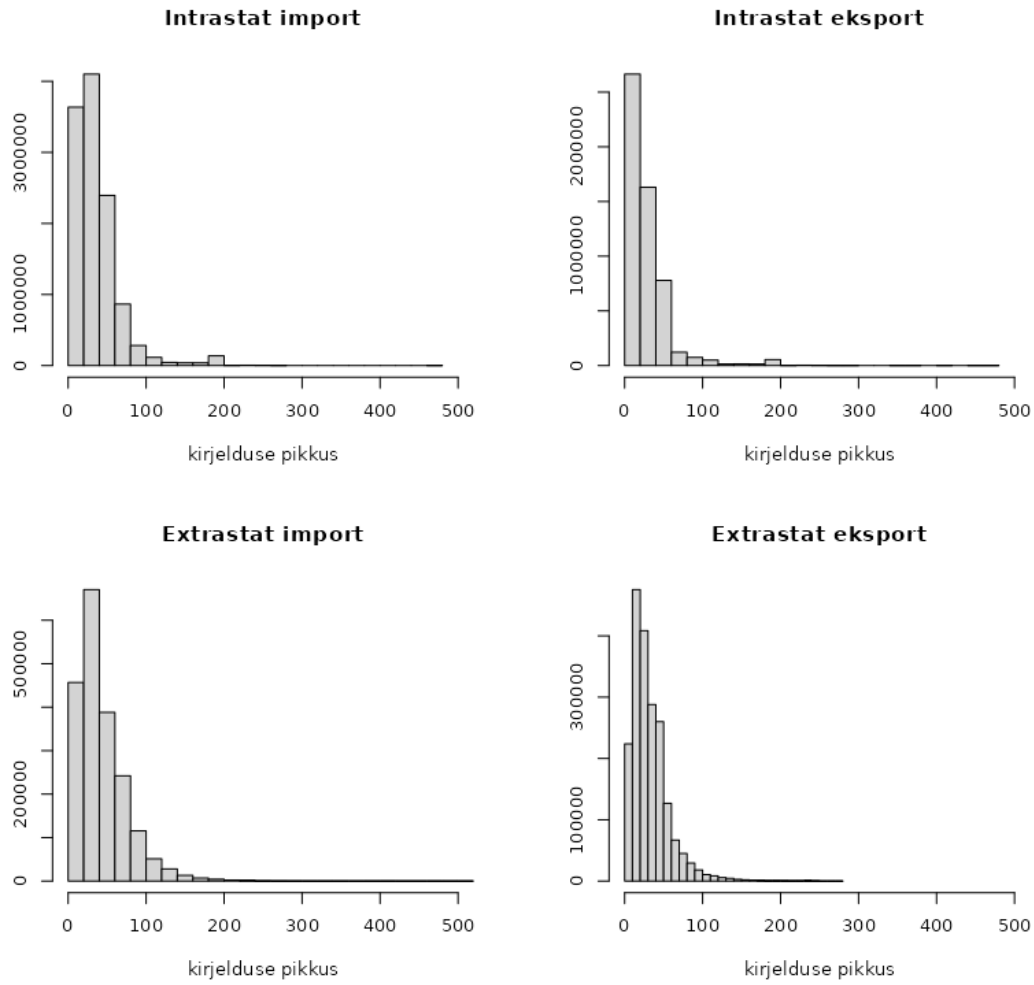



Kaupade kirjeldused on enamasti pigem lühikesed (tabel 6, joonis 2). Keskmiselt on kirjeldused pikemad impordivoogudes ja Extrastati uuringus.

**Tabel 6. Kaubakirjelduste pikkus, tähemärke (eemaldatud kirjavahemärgid). Valikus kõik kirjed aastatest 2019–2023, kus kaubakirjeldus on vähemalt 3 tähemärki**

| | Intrastat import | Intrastat eksport | Extrastat import | Extrastat eksport |
|---|---|---|---|---|
| **Mediaan** | 31 | 21 | 35 | 28 |
| **Keskmine** | 36,5 | 28,1 | 44,1 | 33,2 |
| **Standardhälve** | 30,3 | 28,5 | 31,9 | 24,5 |
| **N kirjeid** | 11 661 402 | 5 423 262 | 1 982 058 | 1 982 848 |

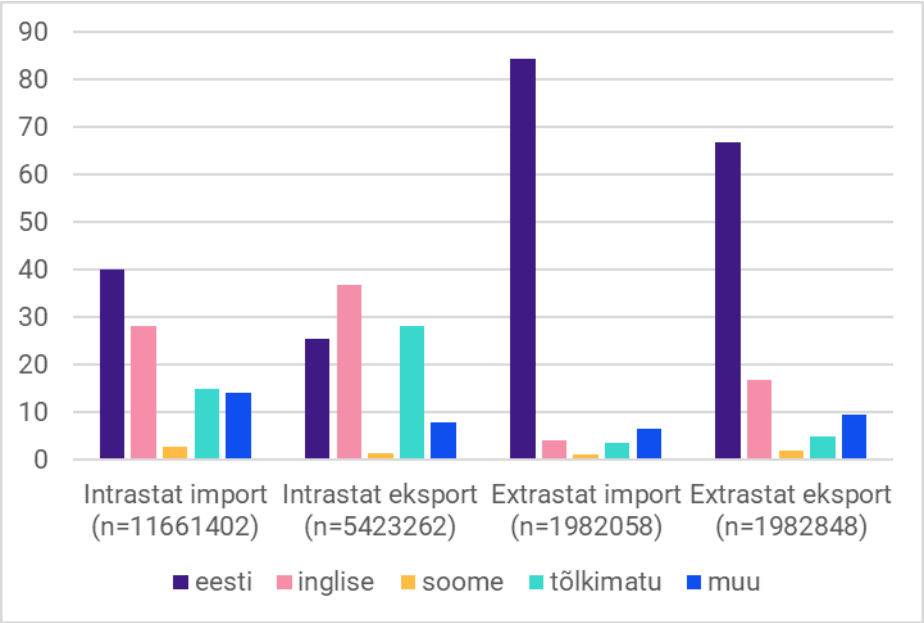
**Joonis 2. Kaubakirjelduste pikkused uuringu ja voo järgi. Valikus kõik kirjed aastatest 2019–2023, kus kaubakirjeldus on vähemalt 3 tähemärki**




Kaubakirjelduste keelelist jaotust hinnati Google Mediapipe (Lugaresi *et al.*, 2019) tarkvara keeletuvastuse mudeliga[1]. Extrastati andmestiku kaubakirjeldused on suures osas eestikeelsed (joonis 3), kuigi ekspordivoos on ka arvestataval määral ingliskeelseid kirjeldusi. Intrastati puhul on keeleline mitmekesisus märgatavalt suurem, kuigi domineerivad eesti ja inglise keel. Lisaks on Intrastatis, eriti ekspordivoos suur osa kirjeldusi, mida mudel ei suutnud ühegi keele alla klassifitseerida (ilmselt suuresti kaubakoodid või muud sümbolite kombinatsioonid).

[1] https://ai.google.dev/edge/mediapipe/solutions/text/language_detector/python

**Joonis 3. Kaubakirjelduste keeleline jaotus, %**




# Potentsiaalsed andmeallikad

Suurema masinõppeandmestiku moodustamiseks on vaja leida selliseid kaubakoodi ja kirjelduse paare, mis sobivad hästi Eesti Intrastati ja Extrastati uuringutes leiduvate kirjeldustega (keeleline jaotus, kirjelduste pikkus, kaubakoodile omaste märksõnade esinemine), aga mis samas ei ole juba meile kättesaadavates andmetes duplitseeritud. Täiendavate andmete otsimisel tuleks keskenduda just haruldasematele kaupadele, mille kohta kirjeldusi Eestis napib. Sellest lähtuvalt võiks täiendavaid treeningandmeid leida teiste Euroopa riikide väliskaubanduse andmestikest (tolli- ja statistikaametid). Kui Eesti Intrastati uuringu kaubakirjeldustest on kolmandik ingliskeelsed, võib ehk ingliskeelseid haruldaste kaupade kirjeldusi leida üle kogu Euroopa. Eestikeelsete kirjelduste juurde saamiseks ilmselgeid valikuid peale käsitöö ei paista.

# Masinõppe katsed

Kaubakoodi ennustamiseks kirjelduse põhjal tehti väljavõte Intrastat ja Extrastat andmestikest aastatel 2019–2023. Kõigepealt otsiti kirjeldustest juhuslikke sümboleid, mis tekivad koodisüsteemide vahel konverteerimisest andmekogumise käigus, ning asendati need vastavate täpitähtedega. Edasi eemaldati kirjavahemärgid ning seati kogu tekst alaregistrisse. Kuna nende andmestike parandamisel ei ole tuvastatud ja parandatud kõiki kaubakoodi vigu, kasutati andmestiku puhastamiseks *doc2vec* mudelit (Le & Mikolov, 2014; Wijffels, 2022). Selleks eemaldati kõigepealt andmestikust multiplikaate nii, et iga kaubakoodi kohta jääks alles maksimaalselt kuus identset teksti. Kui jätta sisse kõik multiplikaadid, oleks mudel liialt mõjutatud kirjeldustest, mida mõni andmeesitaja iga kuu identselt esitab. Kui aga eemaldada kõik duplikaadid, oleks kaubakoodi veaga tekstijupil mudelile sama suur mõju nagu korrektsetel kirjetel. Jättes alles ainult mõned multiplikaadid, saab *doc2vec* mudel õppida paremini kaubakoodile omast teksti mustrit, samal ajal luues teatavat kontrasti haruldasemate kaubakoodi vigadega (eeldame, et identse kirjeldusega kaubakoodi vigu on palju vähem kui korrektseid identse kirjeldusega ridasid). Järk-järgult treeniti *doc2vec* mudeleid varieerides hüperparameetreid ning kasutati neid algse, kõigi multiplikaatidega andmestiku peal, et ennustada kaubakoodi. Iga iteratsiooniga leiab uus mudel mingi hulga kirjeid, kus ennustatud kaubakood on identne andmebaasis registreeritud kaubakoodiga. Kuna kaubakoodide arv on väga suur, siis on üpris väike tõenäosus, et kaubakoodi veaga kirjele ennustaks mudel täpselt sama valet kaubakoodi. Sedasi destilleeriti usaldusväärseid kirjeid viie iteratsiooniga, kuni ligi 50% algsest andmestikust määratleti usaldusväärseks. Viimases iteratsioonis õpiti *doc2vec* mudel nendelt usaldusväärsetelt kirjetelt ning ennustati viimane ring kaubakoode. Lõpuks jäi alles 12,2 miljonit usaldusväärset kaubakoodi ja kirjelduse paari, millest moodustati treeningandmestik fastText (Bojanowski *et al*., 2017; Mouselimis, 2024) teksti klassifitseerimise mudeli treenimiseks.

Katsetel valiti puhastatud andmestikust kõik kaubakoodid, millel on vähemalt 30 unikaalset minimaalselt kuue märgi pikkust kirjeldust andmestikus. Võrdluseks võeti samad kaubakoodid puhastamata andmestikust (Tabelis 7 „mustad“ mudelid). Tehti ka eraldi mudel Extrastati andmestiku kohta, kus on kauba kirjeldus kohustuslik. Klasside ebaühtlase esindatuse vähendamiseks valiti mõlemast andmestikust üle 300 unikaalse kirjeldusega kaubakoodide puhul juhuslikult 300 kirjeldust. Enamikes katsetes valiti andmestikust juhuslikult 80% treenimiseks ning 20% testimiseks. Extrastati mudelit testiti ka tasakaalustatud testandmestikuga, kus on 10 kirjet igast kaubakoodist. Kõiki mudeleid treeniti samade hüperparameetritega[2].

[2] Sõnaosa minimaalne pikkus (`minn`) 2, maksimaalne (`maxn`) 5; sõna n-gramme (`wordNgrams`) 1; dimensioone (`dim`) 120; kontekstiakna suurus (`ws`) 3; õppimismäär (`lr`) 0.2; treenimise epohhe (`epoch`) 30.

**Tabel 7. FastText mudelid kaupade klassifitseerimiseks tekstist**

| Andmestik | Andmestiku suurus | Klasside arv | Kulu | Treening täpsus | Test täpsus | Test makro F1 |
|---|---|---|---|---|---|---|
| **Must numbriteta** | 573 473 | 2586 | 2,66 | 0,765 | 0,56 | 0,582 |
| **Puhas numbriteta*** | 345 708 | 2586 | 1,78 | 0,922 | 0,749 | 0,718 |
| **Must numbritega** | 663 390 | 2892 | 2,25 | 0,823 | 0,621 | 0,636 |
| **Puhas numbritega**** | 418 399 | 2892 | 1,64 | 0,943 | 0,797 | 0,765 |
| **Must, viimane** | 789 123 | 3903 | 2,46 | 0,814 | 0,54 | 0,56 |
| **Puhas, viimane***** | 614 519 | 3903 | 1,63 | 0,926 | 0,76 | 0,73 |
| **Extrastat puhas** | 534 566 | 2910 | 1,5 | 0,932 | 0,777 | 0,767 |
| **Extrastat puhas, tasa-kaalustatud****** | 541 537 | 2910 | 1,36 | 0,935 | 0,757 | 0,753 |

* Algne katse esimese puhastamise iteratsiooni andmetega eemaldades kirjeldustest numbrid.
** Sama esimese puhastamise iteratsiooni andmed, aga säilitades tekstides numbrid.

*** Lõpliku puhastatud andmestiku tulemus.

**** Testandmestikus 10 kirjeldust igast kaubakoodist.

Tabelist 7 saab esiteks järeldada, et kaubakirjeldustes sisalduvad numbrid kannavad olulist eristavat infot kaupade klassifitseerimiseks. Mõnedel juhtudel on ilmselt juba kirjelduses sees õige kaubakood. Teiseks saame järeldada, et andmete puhastamise strateegia töötab hästi ning usaldusväärseid kirjeid saaks ilmselt destilleerida veelgi rohkem. Teistpidi viitab puhastamata ja puhastatud andmetikelt saavutatav erinev täpsus arvestatavale kaubakoodi vigade määrale Intrastati ja Extrastati uuringutes. Kolmandaks näib, et võrreldes kirjanduse ülevaates paistvate tulemustega, saavutab *fastText* (kiirusele, mitte maksimaalsele täpsusele optimeeritud algoritm) 3903-klassilise ülesande puhul muljetavaldava täpsuse. Seda saab ilmselt ühtpidi seletada meile kättesaadava andmestiku suurusega – kirjanduse ülevaates kajastuvates uuringutes ei ole autoritele kättesaadav kogu kohaliku statistikaameti andmevara. Teistpidi võib üllatavalt kõrget täpsust seletada *doc2vec* iteratiivse puhastamise tehnikaga – see meetod ilmselt jätabki sõelale lihtsamini eristatavaid kaubakoode, mitte ilmtingimata kõige sagedamini esinevaid kaubakoode. 3903 kaubakoodi kõige suuremast mudelist moodustavad 92% kogu kombineeritud andmestikust ning 2910 kaupa Extrastati mudelist moodustavad 92,4% kogu Extrastati andmestikust (ilma koodita 99510000). Selliste mudelite täisautomaatne kasutamine seega määraks ~8% haruldasematele kirjeldusega kaupadele süstemaatiliselt vale koodi nende sagedasemate koodide seast.

## Masinõppe ülesande keerukus

Masinõppe ülesandeks on jagada kaubakirjelduste tunnusruum tuhandeteks eristatavateks osadeks. Tegemist on kolossaalse ülesandega, mida komplitseerivad veel mitmed asjaolud. Tabelis 8 on välja toodud peamised probleemid ning võimalikud leevendused.

**Tabel 8. Masinõpet komplitseerivad asjaolud ning potentsiaalsed lahendused**

| Probleem | Leevendus |
|---|---|
| **Suur klasside arv (CN4 tasemel sadu, CN6 ja CN8 tasemel tuhandeid)** | • Hierarhiline mudelite süsteem: ennusta kõigepealt laiemat HS2 kaubagruppi ning siis selle järgi vali kitsam mudel, mis on treenitud vastava kaubagrupi CN8 koode eristama (lõplik täpsus sõltuks siis mitme mudeli täpsustest, seega ei pruugi nii väga head tulemust ikkagi saavutada) |
| **Klasside ebavõrdne esindatus / haruldased kaubad** | • Andmestiku moodustamisel rakendada eraldi strateegiat haruldastele kaubakoodidele, näiteks valida kaubakirjeldusi kaugemalt minevikust.<br>• Sagedastel kaupadel näha rohkem vaeva andmete puhastamisega (kaubakoodi vigade eemaldamine) |
| **Kirjeldused on tihti liiga lühikesed / väheinformatiivsed*** | • Kasutada lisainformatsioonina kilohinda (teksti mudeli ning ühe muutujaga arvulise mudeli fusioon ei pruugi olla tehniliselt lihtne ning kilohinna lisainfo ei pruugi olla piisav)<br>• Motiveerida andmeesitajaid pikemaid ja põhjalikumaid kirjeldusi kirjutama (tõenäoliselt ei ole realistlik) |
| **Müra kirjeldustes**** | • Arendada välja regulaaravaldistel põhinev teksti puhastamise süsteem tüüpilisemate mürajuhtumite puhastamiseks (juhusliku müra ja haruldaste trükivigade jaoks ilmselt head lahendust ei ole) |
| **Kaubakirjeldused mitmes keeles, eesti keel on väike keel, mille automaatne tõlge ei ole kõige täpsem.** | • Kakskeelsed mudelid: kõik eestikeelsed sõnad tõlkida inglise keelde ning ühendada algse eestikeelse kirjeldusega ning vastupidi (kogu süsteemi täpsus sõltuks siis keeletuvastuse- ja tõlkimismudelite täpsusest ning kakskeelne mudel ei lahendaks kõigi teiste keelte probleemi) |
| **Märgendi kvaliteet / kaubakoodi vead** | • Kaubakoodi vigade eemaldamine treeningandmestikust keeletehnoloogiate abil<br>• Testandmestiku puhul mingi osa käsitsi üle vaadata (töömahukas ja tüütu) |

* 21 miljoni kirjelduse väljavõttes esineb 480 536 kirjeldust „eur“ ning 83 676 kirjeldust „kaup“.

** Juhuslikud sümbolid (näiteks „mähis“ asemel on „m„his“), trüki- ja kirjavead.

## Kasuteguri hindamine

Võttes arvesse kirjanduse ülevaadet, masinõppe katseid ning olemasolevate andmete analüüsi, saab prognoosida potentsiaalse arendatava süsteemi maksimaalset võimalikku kasutegurit. Kaalutakse kaht tüüpi rakendusi: täisautomaatne kaupade klassifitseerimine (eeldab suurt top-1 täpsust) ning interaktiivne tööriist, kus mudel pakub operaatorile arvu n võimalikku kaubakoodi (eeldab suurt top-n täpsust võimalikult väikese n väärtuse juures).

Kaubakoodi ennustamise mudeli kasutegur r (näitab, kui suurele osale andmestikust suudaks väljaarendatud süsteem õiget kaubakoodi pakkuda)

$$r = a \times b \times c, \text{ kus}$$

$a =$ Kvaliteetsete kirjelduste osakaal $=$ kirjelduste osakaal, mille pikkus on üle 6 sümboli (0,452 kombineeritud andmestikus, 0,992 Extrastatis) × (1- ebainformatiivsete kirjelduste osakaal) $(0{,}95^3) = 0{,}429$ kombineeritud andmestikus, 0,942 Extrastatis;

$b = 1 -$ mudelist kõrvale jäetavate haruldaste kaupade osakaal minimaalselt kuue tähemärgiga kirjeldusega kirjetest $= 0{,}966$ kombineeritud andmestikus (arvestatud 3903 masinõppe katse kaupa), 0,924 Extrastatis (2910 kaupa);

$c =$ Testandmestikult mõõdetav täpsus: kirjanduse ülevaate põhjal 0,46 (Chen *et al*., 2021), suuruse n piisavalt suure väärtuse puhul võimalik saavutada 0,95 topp-n täpsus.

## Tabel 9. Võimalik saavutatav kasutegur ning selle komponendid eri kontekstides

| Stsenaarium | $a$ | $b$ | $c$ | Kasutegur $r$ |
|---|---|---|---|---|
| **Top-1 täpsus, kaasatud kõik kaubad*** | 0,429 | 1 | 0,46 | 0,197 |
| **Top-1 täpsus, kaasatud kaubad, millel on min. 100 unikaalset kirjeldust**** | 0,429 | 0,883 | 0,717 | 0,271 |
| **Top-1 täpsus, kombineeritud andmestik (min. 30 krjeldust)** | 0,429 | 0,966 | 0,76 | 0,315 |
| **95% top-13*** täpsus, kombineeritud andmestik** | 0,429 | 0,966 | 0,95 | 0,394 |
| **Top-1 täpsus, Extrastat tasakaalustamata** | 0,942 | 0,924 | 0,777 | 0,676 |
| **Top-1 täpsus, Extrastat tasakaalustatud** | 0,942 | 0,924 | 0,757 | 0,659 |
| **95% top-9 täpsus, Extrastat** | 0,942 | 0,924 | 0,95 | 0,827 |

* Eeldame Chen *et al*. (2021) täpsust.

** Analoogne ülesanne nagu kirjeldab Paramartha *et al*. (2021), seega täpsuse hinnang sellest artiklist.

*** Operaatorile pakutavate potentsiaalsete kaubakoodide arv, mille seast tuleb valida.

[3] Pealiskaudne hinnang. Tegelikult võib klassifitseerimiseks ebapiisava infoga kirjeldusi olla üle 5%.

Näib (tabel 9), et praegu kättesaadava andmestikuga saab teha täisautomaatse süsteemi, mis suudaks määrata õige kaubakoodi 66%-le tehingutest kolmandate riikidega. Interaktiivne tööriist, mis pakuks 9 kõige tõenäolisemat koodi, suudaks kasutaja abiga määrata õige kaubakoodi 83%-le Extrastati kirjetest. Suurema ja kvaliteetsema treeningandmestikuga on ehk võimalik jõuda 90% lähedale. Peamiselt puuduvate kirjelduste tõttu ei ole aktsepteeritava kasuteguri saavutamine kombineeritud andmestikus ega Intrastatis võimalik.

## Rakendatavus teistes valdkondades

Kombineeritud nomenklatuur on väga suur klassifikaator ning klassifitseeritavad tekstid on andmete olemuse tõttu paljudes eri keeltes. Selget kolmandat rakenduskohta peale tolli ja väliskaubanduse statistika sellele mudelile ei paista. Treenides aga mudelit ainult eestikeelsete kaubakirjeldustega, võib kaubakoodi prognoosimise mudel leida kasutust näiteks eestikeelsetest veebipoodidest kraabitavatel andmetel (võimalik kasutus tarbijahinnaindeksis). *FastText*-algoritmist võib üldiselt olla kasu teiste klassifikaatorite automatiseerimisel, kus leidub tekstilisi kirjeldusi (nt ameti klassifitseerimine töökuulutuse teksti järgi). Andmestiku puhastamiseks kasutatud meetod *doc2vec*-algoritmiga töötab seda paremini, mida rohkem erinevaid klasse masinõppe ülesandes on. Väiksemate klassifikaatorite automatiseerimisel sellest meetodist nii palju kasu ei ole (mida vähem klasse, seda suurem tõenäosus, et *doc2vec* ennustab märgendi veaga kirjele selle sama vale märgendi).

## Kokkuvõte

Rakendusuuringu eesmärgiks oli välja selgitada, kas ja kuidas automatiseerida kaupade klassifitseerimist kombineeritud nomenklatuuri kategooriatesse kaubakirjelduste põhjal ning milliseid tulemusi see võiks anda. Selgub, et tegu on äärmiselt keerulise masinõppe ülesandega, mida omakorda piirab andmekvaliteet (kaubakoodi vead treeningandmetes, puuduvad, mürased ja väheinformatiivsed kirjeldused). Rakendusuuringust võib järeldada, et kaubakoodi klassifitseerimise mudel saab praktilist kasu tuua vaid kaupade deklareerimisel tollis, kus kauba kirjelduse esitamine on kohustuslik. Sellel juhul tuleb siiski arvestada, et haruldastele, treeningandmestikust kõrvale jäänud kaupadele (~8% juhtudest Extrastati mudeli puhul) ennustaks mudel alati valet koodi ning kodeerija peaks ikkagi nomenklatuuri leksikoni poole pöörduma (eeldusel, et ta saab aru, et pakutud 9 koodi seas õiget koodi ei ole). Euroopa-siseste tehingute (Intrastat) puhul on suurem keeleline mitmekesisus ning kaubakirjeldused lühemad, kui neid üldse on – praktilist kasu andmestiku, mudeli ja rakenduse välja arendamisest palju ei saaks.

# Kasutatud kirjandus


Altaheri, F., & Shaalan, K. (2020). Exploring machine learning models to predict harmonized system code. In *Information Systems: 16th European, Mediterranean, and Middle Eastern Conference, EMCIS* 2019, Dubai, United Arab Emirates, December 9–10, 2019, Proceedings 16 (pp. 291–303). Springer International Publishing.

Binh, N. T., Nguyen, H. A., Linh, P. N., Giang, N. L., & Thang, T. N. (2021). Attentive RNN for HS Code Hierarchy Classification on Vietnamese Goods Declaration. In *Intelligent Systems and Networks: Selected Articles from ICISN 2021*, Vietnam (pp. 298–304). Springer Singapore.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers, 427–431.

Chen, X., Bromuri, S., & Van Eekelen, M. (2021, April). Neural machine translation for harmonized system codes prediction. In *Proceedings of the 2021 6th International Conference on Machine Learning Technologies* (pp. 158–163).

Ding, L., Fan, Z., & Chen, D. (2015). Auto-categorization of HS code using background net approach. *Procedia Computer Science*, 60, 1462–1471.

Du, S., Wu, Z., Wan, H., & Lin, Y. (2021, May). HScodeNet: Combining hierarchical sequential and global spatial information of text for commodity HS code classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 676–689). Cham: Springer International Publishing.

He, M., Wang, X., Zou, C., Dai, B., & Jin, L. (2021). A commodity classification framework based on machine learning for analysis of trade declaration. *Symmetry*, 13(6), 964.

Jahanshahi, H., Ozyegen, O., Cevik, M., Bulut, B., Yigit, D., Gonen, F. F., & Başar, A. (2021, November). Text classification for predicting multi-level product categories. In *Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering* (pp. 33–42).

Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196). PMLR.

Lee, E., Kim, S., Kim, S., Jung, S., Kim, H., & Cha, M. (2024). Explainable Product Classification for Customs. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–24.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

Mouselimis, L. (2024). fastText: Efficient Learning of Word Representations and Sentence Classification using R. R package version 1.0.4, https://CRAN.R-project.org/package=fastText.

Paramartha, I. G. Y., Ardiyanto, I., & Hidayat, R. (2021, April). Developing machine learning framework to classify harmonized system code. case study: Indonesian customs. In *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)* (pp. 254–259). IEEE.

Wijffels, J. (2022). doc2vec: Distributed Representations of Sentences, Documents and Topics. R package version 0.2.0, https://CRAN.R-project.org/package=doc2vec.