UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Karmen Kink

Classification of E-Commerce Products Based on Textual Product Descriptions

Master's Thesis (30 ECTS)

Supervisors: Kairit Sirts, Karl-Oskar Masing

Tartu 2021

Classification of E-Commerce Products Based on Textual Product Descriptions

Abstract:

Assigning Harmonized System (HS) codes to products is necessary to comply with customs regulations, gather statistics, and prevent fraud. Since the HS is a complex system with many classes, automatic HS code classification is required to speed up the process and ensure correctness. In this thesis, we explore two types of machine learning methods for HS classification – shallow neural network classifiers and deep neural network classifiers that are based on the Transformer architecture. We find that with a large dataset, shallow classifiers can be relatively easily improved to outperform Transformer-based classifiers, while tuning the latter is a more complex and time-consuming task. We also discover that the training data includes erroneously labeled entries and that this has a negative impact on the models.

Keywords:

Natural language processing, Harmonized System, multi-class classification

CERCS:

P176 Artificial Intelligence

E-kaubanduse toodete klassifitseerimine tekstiliste tootekirjelduste alusel

Lühikokkuvõte:

Harmoneeritud Süsteemi (HS) koodide määramine toodetele on vajalik tollinõuetele vastamiseks, statistika kogumiseks ja maksupettuste vältimiseks. Automaatne HS-koodide klassifitseerimine aitab aega säästa, sest HS on kompleksne ja paljude klassidega süsteem, mistõttu käsitsi õige klassi valimine on keerukas ja aeganõudev. Selles magistritöös kasutame HS-koodide klassifitseerimiseks kaht tüüpi masinõppemeetodeid: pindmistel tehisnärvivõrkudel põhinevaid klassifitseerijaid ja sügavatel tehisnärvivõrkudel, täpsemalt Transformer-arhitektuuril, põhinevaid klassifitseerijaid. Selgub, et sellise suurusega andmestiku puhul, nagu meil on võimalik kasutada, saab pindmisi klassifitseerijaid võrdlemisi kergesti arendada maani, kus nad annavad paremaid tulemusi kui sügavad närvivõrgud, viimaseid on aga oluliselt keerukam ja aeganõudvam edasi arendada. Lisaks leidsime, et kasutatav andmestik sisaldab valesti märgendatud kirjeid ning et sellel on negatiivne mõju mudelite kvaliteedile.

Võtmesõnad:

Loomuliku keele töötlus, Harmoneeritud Süsteem, mitmeklassiline klassifitseerimine

CERCS:

P176 Tehisintellekt

License

Non-exclusive licence to reproduce thesis and make thesis public

I, Karmen Kink,

herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from ***14/05/2022*** until the expiry of the term of copyright,

Classification of E-Commerce Products Based on Textual Product Descriptions,

supervised by Kairit Sirts and Karl-Oskar Masing,

2. I am aware of the fact that the author retains the rights specified in p. 1.

3. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Karmen Kink

14/05/2021