COICOP flokkari

Bergur Þorgeirsson
Hugbúnaðarsérfræðingur hjá Hagstofu Íslands
bergurth@hi.is

Indriði Arnaldsson
Meistaranemi
ina23@hi.is

Abstract

Í þessu verkefni bjuggum við til flokkara fyrir COICOP gögn. Við beyttum einna helst þremur reikniritum við gerð flokkarans og munum fara yfir niðurstöðurnar í þessari skýrslu.

1 Inngangur

Gögn frá matvöruverslanakeðjum rata til Hagstofunnar með vörulýsingum, en gagnlegt getur verið að flokka þessar vörulýsingar í flokka út frá COICOP2018 staðlinum(United Nations , 2018). Þessi flokkun er meðal annars notuð í undirbúningi reiknings á vísitölu neysluverðs. Þetta ferli er í einhverjum mæli handvirkt núna, einkum þegar um er að ræða nýjar vörulýsingar sem ekki hafa verið flokkaðar áður, og því er möguleiki á hagnýtingu afurða verkefnisins í því skyni að spara vinnu í framtíðinni. Fram hafa farið áður ýmsar tilraunir til sambærilegrar sjálfvirkrar flokkunar hjá Hagstofunni. Til dæmis var niðurstöðum skilað úr verkefni styrkt af Eurostat árið 2018(Eurostat, 2018), en þar voru prófaðar tvær aðferðir. Annars vegar svo kallað "Fuzzy matching"sem notar Levenshtein distance og hins vegar vélanáms aðferð "Decision Forest/jungle". Í stuttu máli náðist á þeim tíma accuracy upp á 64% úr fuzzy matching aðferðinni en accuracy up á 54% úr random decision forest aðferðinni.

Við ákváðum í þessu verkefni að athuga eina einfalda tölfræðilega flokkunaraðferð, Naive-bayes, og studdumst við útfærslu úr sklearn python pakkanum, en einnig vildum við prófa að fínþjálfa flokkara út frá nokkrum almennum mállíkönum, sem við notuðum sem grunnlíkön til að fínþjálfa flokkara okkar ofan á. Þessi líkön voru: XLMR-ENIS líkan sem er tvímála transformerlíkan þjálfað á íslenskum og enskum textum https://huggingface.co/vesteinn/XLMR-ENIS, IceBERT sem er íslenskt transformerlíkan þjálfað á íslenskum textum ofan á RoBERTa-base https://huggingface.co/vesteinn/IceBERT, einnig kom til greina að þjálfa ofan á fastText_is_rmh mállíkaninu https://huggingface.co/vesteinn/fasttext_is_rmh sem og ofan á opna fastText líkaninu frá Facebook's AI Research https://fasttext.cc. Okkur fannst áhugavert að sjá hvernig notkun mállíkana mundi berast saman við naive-bayes og líka spennandi að sjá hvaða líkön væru að skila bestu niðurstöðunum.

2 Gagnasöfnun

Við fengum aðgang að þrem dálkum úr töflu scanner-data gagna sem búið var að flokka handvirkt í COICOP flokka. Dálkarnir voru 'Heading', 'coicop2018' og 'vorulysing'. Dálkurinn 'coicop2018' hafði að geyma flokkunina en 'Heading' í raun bara lýsingu á þeim flokki á meðan að 'vorulysing' hafði að geyma inntakið sem hafði verið flokkað í viðkomandi flokk. Um var að ræða 20142 línur af lýsingum, flokk og heading.

Eftir að innihald þessara dálka var komið yfir í eina stóra csv skrá, gerðum við lítið forrit sem tók öll gögnin og skipti þeim slembið en jafnt á alla flokka yfir á þrjú ný skjöl.Þessi skipting var 80% fyrir þjálfunargögn(e. train set) svo 10% fyrir prófunargögn(e. test set) svo aftur 10% fyrir þróunargögn(e. validation set). Hér fyrir neðan má sjá dreifingu fjölda lína á hvern flokk, en eins og sjá má er dreifingin frekar ójöfn.

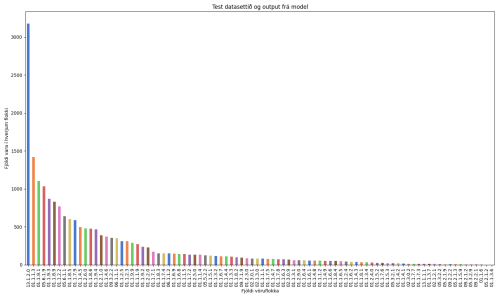

Figure 1: Dreifing gagna yfir alla COICOP flokka

þegar búið var að skipta línunum upp í þjálfunar, prófunar og þróunar gögn, þannig að hlutfall flokka væri sambærilegt innan hvers gagnasets, þá var hvert gagnasett vistað á bæði csv og json formi, eins var skipt út nöfnum allra flokka fyrir tölur frá 0 til 90, en haldið var utan um venslin á milli nýju flokksnafnanna og upprunalegu coicop2018 flokkanna í json skrá sem fékk heitið cat_mapping.json. Sjá má skriptuna sem flokkaði gögnin hér. https://github.com/Bergurth/coicop_text_categorization/blob/master/dataPrep.py

3 Aðferðafræði

Þjálfuð voru fimm mismunandi líkön á meðan á verkefninu stóð; Naive-Bayes, XLMR-ENIS, IceBERT, fastText, og íslensk útfærsla af fastText. Um var að ræða þrjú mismunandi form á inntaks gögnunum, sömu gögnunum fyrir hin mismunandi líkön. FastText líkönin tóku við gögnum á einu ákveðnu formi(Joulin, A, o.fl., 2016), Naive-Bayes á öðru og þjálfunar forrit okkar fyrir XLMR-ENIS og IceBERT tóku við gögnunum á en öðru formi. Eftir að líkönin voru þjálfuð þurfti að ganga frá þeim þannig að þau gætu skilað úttaki sínu á sömu formi m.t.t. hvors annars, til að auðvelda samanburð. Stuðst var við classification_report fallið úr sklearn.metrics pakkanum til að bera saman líkönin.

Við þjálfun á IceBERT og XLMR-ENIS líkönum var notast við transformers python pakkann og unnið var út frá kennslu verkefni https://colab.research.google.com/github/huggingface/notebooks/blob/master/examples/text_classification.ipynb sem við aðlöguðum að okkar þörfum, einkum m.t.t. þess að flokka í fleiri flokka. Hægt er að sjá lýsingu á þjálfunarskrefum eins og þau voru fyrir XLMR-ENIS og IceBERT líkönin hér: https://github.com/Bergurth/coicop_text_categorization/blob/master/trainBERT-like-coicop.py . Taka ber fram að notað var stilling load_best_model_at_end = True, sem veldur því að checkpoint fyrir epoch-ið sem stóð sig best á þjálfunar-metric-inu er svo valin í lok þjálfunar. Með notkun skjákortsins "NVIDIA Tesla P100 GPU"var sá tími sem fór í að keyra circa 20 epoch af þjálfun u.þ.b. 20 til 30 mínútur. Stillingar(e. hyperperameters) sem voru notaðar við þjálfunina voru þær sömu og voru fyrir í kennsluverkefninu sem við studdumst við, en ekki gafst tími fyrir okkur að komast að því nákvæmlega hvaða hyperparameters kynnu að vera bestir.

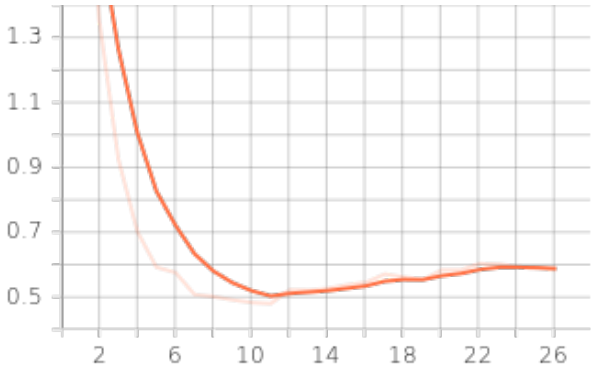

Figure 2: Loss m.t.t. prófgagna eftir epoch

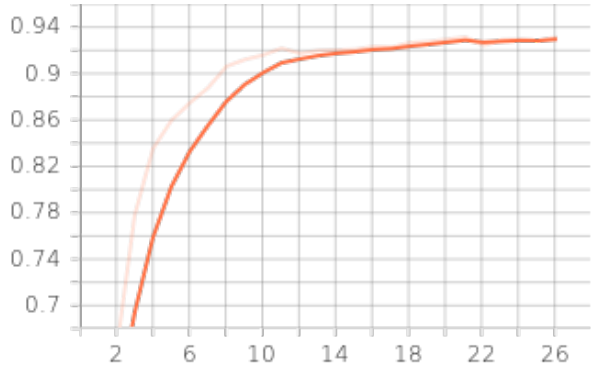

Figure 3: Accuracy m.t.t. prófgagna eftir epoch

Naive-bayes aðferðin sem við notuðum var útfærsla af multinomial Naive-bayes frá sklearn python pakkanum, en multinomial er útvíkkun á binomial dreifingu og hentar þegar um svona flokka er að ræða. Þjálfun þess tók mjög stuttan tíma og sérstaklega í samanburði við hin reikniritin sem við notuðum.

Þjálfun fastText líkananna tók u.þ.b. 5 mínútur á venjulegu CPU örgjörva.

4 Niðurstöður og prófanir

Til þess að meta getu hina ýmsu líkana var stuðst við classification_report fall úr python sklearn.metrics pakkanum. Viðkomandi fall tekur inn lista af sönnu Y annars vegar og hins vegar lista af Y gildum sem viðkomandi líkan hefur giskað á. Fallið gefur mjög ýtarlega greiningu niður á hvern flokk á precision recall og f1. Hér fyrir neðan sýnum við aðeins meðaltölin. Fyrir lengri flokkunarskýrslur fyrir hvert líkan sjáið hér https://github.com/Bergurth/coicop_text_categorization/tree/master/metric_reports

model	precision	recall	f1-score	support
XLMR-ENIS				
accuracy			0.93	2059
macro avg	0.86	0.84	0.84	2059
weighted avg	0.93	0.93	0.93	2059

model	precision	recall	f1-score	support
IceBERT				
accuracy			0.91	2059
macro avg	0.81	0.80	0.79	2059
weighted avg	0.91	0.91	0.91	2059

model	precision	recall	f1-score	support
fastText				
accuracy			0.91	2059
macro avg	0.80	0.79	0.79	2059
weighted avg	0.91	0.91	0.91	2059

model	precision	recall	f1-score	support
isl-fasttext				
accuracy			0.87	2059
macro avg	0.80	0.77	0.78	2059
weighted avg	0.87	0.87	0.87	2059

model	precision	recall	f1-score	support
naive-bayes				
accuracy			0.77	2059
macro avg	0.61	0.41	0.46	2059
weighted avg	0.77	0.77	0.74	2059

XLMR-ENIS líkanið var þjálfað í 26 epoch, sjálfvirkt var epoch 21 valið þar sem það stóð sig best með 93% accuracy, . Var það líkan þjálfað m.t.t. accuracy yfir þjálfunargögnin.

Í tilfelli IceBERT var þjálfunin keyrð í 26 epoch en epoch 18 var valið sjálfvirkt þar sem það stóð sig best. Besta accuracy sem IceBERT náði var 91%.

FastText módelin bæði voru þjálfuð með hjálp "autotuneValidationFile"perameter, sem með því að vísa í prófunargögn var hægt að sjálfvirkt stilla hyperparameters til að ná sem bestum árangri. Var haft sérstaklega fyrir því að koma íslenska fastText líkaninu yfir á facebook fastText form í því skyni að geta nýtt þennan eiginleika. En enska fastText stóð sig betur með 91% accuracy á meðan íslenska náði 87%.

Sjá má að macro meðaltal fyrir recall og percision eru áberandi slæm í naive-bayes. Macro meðaltalið gengur út frá því að allir flokkar eru vigtaðir jafnt óháð hversu algengt sé að vara komi fyrir í þeim flokk. Þessar tölur þ.e. macro average fyrir percision og recall gefa svolítið til kynna hvernig upplifun er af því að nota viðkomandi líkan, ef gengið er út frá því að maður er með nýja vörulýsingu sem hefur ekki komið fyrir áður og maður vill að hún flokkist vel nokkuð óháð því hversu algengt er að vara komi fyrir í rétta flokk viðkomandi lýsingar.

5 Lokaorð og næstu skref

Verkefnið gekk vonum framar og það er ekki hægt að segja að við hefðum búist við svona niðurstöðum þegar við vorum nýbyrjaðir. Það er erfitt að segja til um næstu skref. Eins og stendur er ekki áætlað að við munum halda áfram með verkefnið. Það væri hægt að gera meira, bæði að prufa að keyra gögnin í gegnum fleiri reiknirit og síðan væri einnig hægt að finna mögulega betri stillingar (e. hyperperameters) fyrir þjálfun á einhverjum af þeim líkönum sem við notuðum. Ef ske kynni að eitthvað af líkönunum

yrðu nýtt, eða færð í rekstur á einhvern hátt, þá væri ekki verra að hafa það í huga að flokka reglulega handvirkt nýjar lýsingar til að fá reglulega "gold standard"próf gögn flokkuð af fólki til að geta reglulega endurmetið metrics á þeim líkönum sem væru í notkun. Núverandi líkön standa sig best á núverandi þjálfunargögnum og stundum ágætlega á prófgögnunum okkar, en það þýðir ekki endilega að líkönin muni standa sig jafn vel í því að flokka allar mögulegar vörulýsingar sem kynnu að rata til Hagstofunnar í framtíðinni.

Svo mætti kannski nefna það að þessi líkön eru misstór. Naive-Bayes flokkarinn okkar vistast í 116K skrá, fastText líkanið okkar 631M skrá, en til samanburðar tekur eitt XLMR-ENIS líkanið sem við þjálfuðum upp heil 18G.

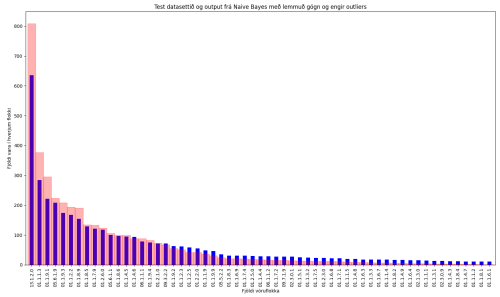

Figure 5: Dreifing á Naive Bayes úttaki miðað við raunveruleg test gögn

6 Töflur og myndir

Hér í gröfum fyrir neðan má sjá dreyfingu vörulýsinga úr prófgögnum í flokka (bláu súlurnar) annarsvegar, og hinsvegar dreyfingu ágiskanna flokkaranna í flokka (gegnsæu bleiku súlurnar). Síðasta grafið er samanburður á öllum líkönum sem við prufuðum.

Ber að hafa í huga að hér er einungis verið að sýna í hversu miklu magni líköninn flokkuðu lýsingar í flokka óháð því hvort sú flokkun sé rétt eða ekki. Graf fyrir IceBERT er ekki sýnt þar sem það líkist mjög mikið grafinu fyrir XLMR-ENIS, og bætir því engu við að sýna það líka.

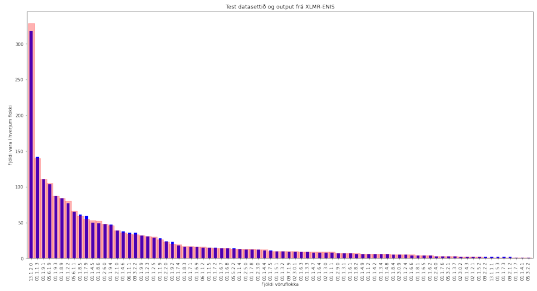

Figure 6: Dreifing á XLMR-ENIS úttaki miðað við raunveruleg test gögn

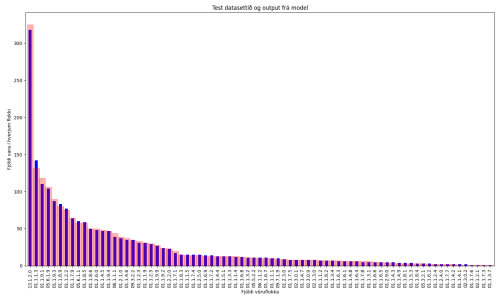

Figure 4: Dreifing á fastText úttaki miðað við raunveruleg test gögn

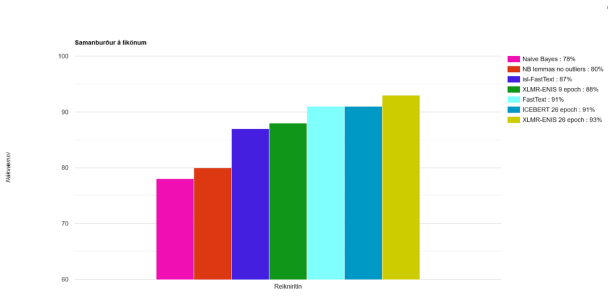

Figure 7: Accuracy allra reiknirita sem við prufuðum borið saman.

Þakkir

Við viljum byrja á því að þakka sérstaklega Vésteini Snæbjarnarsyni fyrir mikla hjálp í öllum ferlum verkefnisins. Einnig viljum við þakka starfsmönnum Hagstofu Íslands þeirra á meðal:
Heiðrún Erika Guðmundsdóttir, Sigurjón Leifsson, Ólafur Arnar Þórðarsson, Arndís Vilhjálmsdóttir, Yayoi Shimomura, Auðunn Ragnarsson, Elsa Björk Knútsdóttir, Sigrún Pálsdóttir, Ólafur Jón Björnsson, Ólafur Hjálmarsson auk annarra sem að verkefninu komu.

Heimildir

United Nations. (2018). Classification of Individual Consumption According to Purpose (COICOP).

Eurostat., (2018), Chapter "Objective 1.A: Automatic linking of GTIN codes / shop specific codes to ECOICOP" from "Module (DTM): 04.1.51, Provide macroeconomic accounts and aggregates supplemented by satellite accounts and measures of social performance. Action 1: Price statistics - methodological and practical improvements and developments. Grant Agreement No. 04151.2016.005-2016.617 between Eurostat and Statistics Iceland: Final Report on four objectives"

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

A Viðaukar

Fyrir áhugasama um Naive Bayes flokkun þá bendum við á kafla í Speech and Language Processing. Eftir Daniel Jurafsky og James H. Martin. Kaflann er hægt að finna hér á þessum link
https://web.stanford.edu/~jurafsky/slp3/4.pdf

Fyrir áhugasama um orðvigra, vigramerkingarfræði og greypingar (tækni sem liggur til grundvallar í þeim mállíkönum sem við þjálfuðum) bendum við á sjötta kafla í áðurnefndri bók.
https://web.stanford.edu/~jurafsky/slp3/6.pdf

og á greininna: Efficient Estimation of Word Representations in Vector Space eftir Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean https://arxiv.org/abs/1301.3781

Fyrir áhugasama um Transformer líkön (eins og XLMR-ENIS og IceBERT) bendum við á greinina: Attention Is All You Need
eftir Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
https://arxiv.org/abs/1706.03762