Áfangaskil - Náttúruvernd í fréttum og á þingi

Elías Bjartur Einarsson (ebe19) og Karl James Pestka (kjp3)

Inngangur

Markmið þessa verkefnis má skipta niður í skref (sjá @fig:1). Í upphafi viljum við skapa máltækniflokkara sem getur greint umræðu um náttúruvernd og loftslagsmál í bæði fréttatextum og alþingisskjölum ásamt annars konar gögnum að einhverju leyti. Því næst viljum við geta greint afstöðu (e. sentiment) þess texta, það má bæði skilja sem jákvæðni eða neikvæðni textans eða þá afstöðu hans til loftslagsbreytinga af mannavöldum (hvort viðkomandi geri mikið eða lítið úr málaflokknum). Í þriðja lagi viljum við geta greint hver sé höfundur textans og hvort viðkomandi tali fyrir hönd stjórnmálaflokks. Að endingu, ef ofangreint gengur upp, viljum við geta flokkað nánara eftir einkunnaflokkana Sólarinnar (https://solin2021.is) þegar kemur að markmiðum tengdum náttúruvernd, loftslagsbreytingum og hringrásarhagkerfinu. Í því tilfelli myndum við nota gögn og einkunnir Sólarinnar sem viðmið.

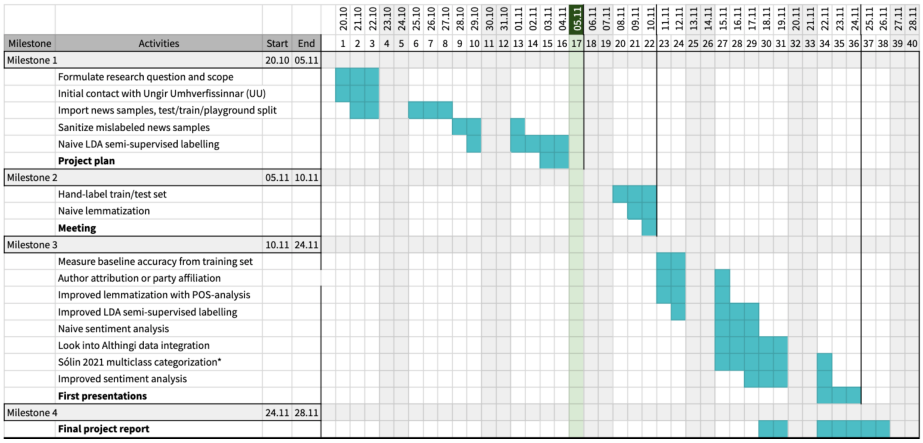

Figure 1: Verkáætlun yfir tíma í GANTT rit.

Gögn

Gögnin sem nú er unnið með eru safn rúmlega 600.000 frétta af íslenskum miðlum, fengnar frá Vésteini. Þeim hefur verið skipt upp í þjálfunarsafn og prófunarsafn, unnið er að leiðum til að merkja hluta þeirra í fréttir tengdar náttúruvernd og fréttir ótengdar þeim málaflokki.

Prófað hefur verið að keyra LDA flokkara á hluta fréttanna sem skiptir þeim í 10 hópa og hefur það gengið vel. Með því væri hægt að plokka út undirsöfn sem annað hvort þykir líklegt að tengist málaflokknum eða þá að tengjast greinilega

ekki flokknum að nokkru leyti. Til stendur að þjálfa stærra líkan sem vinnur með öll þjálfunargögnin.

Önnur gögn sem til stendur að nálgast og vinna með eru annars vegar gögn frá Alþingi og bæjarstjórnum sem og gögn sem Sólin notaði til að setja saman einkunnatöfluna sína. Gögn frá Alþingi og bæjarstjórum verða notuð til að tengja stjórnmálamenn við flokka og greina orð þeirra úr pontunni og frumvörpum. Gögnin Unga Umhverfissinna í verkefninu Sólinni yrðu notuð til að grundvalla nánara málaflokkun líkansins.

Árangur

Hingað til hefur vinnan miðast að því að lesa inn gögnin og fá yfirlit yfir þau, leita í þeim að náttúruverndartengingum og flokka þau með viðgjafarlausri hópun (e. unsupervised clustering) með LDA aðferðum. Hugmyndin er að handmerkja hluta gagnanna, senda gögnin án merkinga í slíka hópun og skoða svo í hvaða hópa jákvætt merktu og neikvætt merktu gögnin lenda. Til þess að slíkt sé marktækt þarf að merkja hluta gagnanna sem prófunar- eða þróunarsafn og athuga hvort aðferðin flokkar þau gögn rétt.

Við höfum einnig eytt tíma í samskipti við Unga Umhverfissinna til að nálgast gögn Sólarinnar en möguleiki er fyrir hendi að þau gögn þyki of viðkvæm til að við getum fengið aðgang að þeim. Við sjáum hvað setur.

Þar að auki höfum við litið til þess að nota flokkara ofan á ICEbert líkan til þess að flokka texta í flokka sem tengjast náttúruvernd og þeir sem ekki tengjast því. Við höfum fundið og niðurhalað Huggingface bók Simple Transformers (sem inniheldur ClassificationModel líkan) í þeim tilgangi.

Næstu skref

Næst á dagskrá er eins og áður segir að merkja gögn handvirkt og koma upp flokkara byggðum á því sem getur með góðu móti flokkað fréttir í þær sem tengjast náttúruvernd og þær sem tengjast því ekki. Þann flokkara munum við bera saman við einfalt grunnlíkan sem við setjum upp. Einnig er á dagskrá að skrapa net Alþingis eða ganga í einhver gagnasöfn sem hafa nú þegar gert það. Við munum halda samskiptum okkar áfram við fulltrúa Sólarinnar í von um aðgang að gögnunum þeirra. LDA flokkarinn sem við notumst við núna notar ansi einfeldningslega lemmun og mögulega er ástæða til að gera það betur á næstu dögum.

Háskóla Íslands, 5. nóvember 2021
Elías Bjartur Einarsson og Karl James Pestka