Diúltach nó Dearfach? Mianadóireacht Tuairime ar Ghiolcacha Gaeilge

Caoilfhionn Lane

https://doi.org/10.13025/2gjc-j517

Réamhrá

Teicneolaíocht Teanga na Gaeilge

Is éard is Teicneolaíocht Teanga ann ná an réimse sin, idir an ríomhaireacht agus an teangeolaíocht, ina ndéanann taighdeoirí anailís agus ionramháil, go minic go huathoibríoch, ar théacs, cibé teanga ina bhfuil sé scríofa. Tá iliomad uirlisí agus modhanna ar fáil don té atá ag obair le teicneolaíocht teanga an Bhéarla, agus dá thoradh seo tá dul chun cinn suntasach déanta i réimsí ar nós an aistriúcháin uathoibríoch, téacs go caint, agus anailís ar thopaicí. Le blianta beaga anuas pléadh go cuimsitheach na dúshláin a bhaineann le teicneolaíocht teanga na Gaeilge (féach, mar shampla, Judge *et al.*, 2012). Tá dul chun cinn ar leith déanta sa réimse seo i nGaeilge, mar shampla san aistriúchán uathoibríoch (Dowling *et al.*, 2015; Arčan *et al.*, 2016), agus san anailís theangeolaíoch de théacs Gaeilge (Lynn *et al.*, 2017). In ainneoin sin, i gcomparáid leis na mórtheangacha, tá an Ghaeilge fós gann in acmhainní ó thaobh teicneolaíochta teanga de, agus dá bharr tá sé níos deacra ag taighdeoirí modhanna teicneolaíochta teanga a chur i bhfeidhm ar théacs Gaeilge.

Mianadóireacht Tuairime

Tá borradh mór faoi réimse ar leith de theicneolaíocht teanga, ar a dtugtar an mhianadóireacht tuairime.[1] Úsáidtear an mhianadóireacht tuairime (tugtar *sentiment analysis* nó *opinion mining* air sa litríocht) chun iniúchadh a dhéanamh ar thuairimí agus mothúcháin atá ag scríbhneoirí téacs. Cuir i gcás go bhfuiltear ag iarraidh dearcthaí nó tuairimí atá ag pobal na meán sóisialta faoi thopaic ar leith a fháil amach. Mar shampla, go cainníochtúil cérbh iad na heachtraí ba mhó a chuaigh i bhfeidhm ar dhaoine ar na meáin shóisialta in 2017? Cérbh iad na topaicí ba dhearfaí nó ba dhiúltaí sa tréimhse chéanna? An féidir na rudaí seo a thomhas? Chun dul i ngleic le ceisteanna mar seo déantar anailís ar na tréithe sin i dtéacs a léiríonn tuairimí, mothúcháin agus dearcthaí (Liu, 2012). Go hiondúil, déantar an anailís seo ar théacs ina bhfuil tuairimí forleathana ann, amhail postálacha ar na meáin shóisialta nó ar léirmheasanna ar líne. Aithnítear go bhfuil focail áirithe ann (*sentiment words* nó focail tuairime[2]) a léiríonn tuairim nó mothúchán ar bhealach níos láidre ná focail eile.

Mar a stóráiltear focail i bhfoclóir, stóráiltear focail tuairime i léacsacan tuairime[3] (*sentiment lexicon*). Sin liosta de na focail tuairime (do theanga ar leith) agus iarracht déanta an tuairim san fhocal a aithint agus, ag brath ar an saghas léacsacain, meáchan nó luach a thabhairt don tuairim sin. Cuir i gcás, tá na focail *álainn* agus *iontach* dearfach, agus tá *olc* agus *uafásach* diúltach. Bíonn focail eile débhríoch ó thaobh tuairime de, agus focail eile neodrach (mar shampla, bíonn *mealltach* débhríoch agus bíonn *triantánach* neodrach). Tá bealaí éagsúla le meáchan a thabhairt do na tuairimí seo.

Tá neart léacsacan tuairime ar fáil i mBéarla. Mar shampla, leis an léacsacan **NRC** (Mohammad agus Turney, 2010), tugtar do gach focal sa léacsacan "luachanna mothúcháin", a roghnaítear as ocht gcinn: *anger, anticipation, disgust, fear, joy, sadness, surprise,* agus *trust.* Tugtar freisin 'luach tuairime' don fhocal – bíonn sé *positive* nó *negative*. Mar shampla, sa léacsacan seo tá na luachanna *fear*, *anger* agus *surprise* ag an bhfocal '*abduction'*, agus is focal *negative* é. Tá na luachanna *joy*, *surprise* agus *trust* ag '*illumination'*, agus is focal *positive* é[4]. Tá samhail léacsacan **Bing** (Hu & Liu, 2004) níos simplí. Tá dhá chineál luach – *negative* nó *positive* – tugtha do gach focal sa léacsacan. Sa léacsacan **AFINN** (Nielsen, 2011), tugtar do gach focal scór idir -5 (an ceann is diúltaí) agus 5 (an ceann is dearfaí).

Tá buntáistí agus míbhuntáistí ag baint leis na léacsacain dhifriúla seo. Tá samhail NRC níos casta agus tá sé deartha don Bhéarla. Bheadh sé deacair é a chur in oiriúint do theanga eile. Ar an lámh eile tá léacsacan Bing simplí ach níos srianta, agus tá sé níos inaistrithe do theangacha eile. Tá go leor léacsacan

1 Is fo-réimse é seo den mhianadóireacht téacs, nó *text-mining* i mBéarla. Sa mhianadóireacht téacs, déantar anailís ar spriocthéacs ionas gur féidir faisnéis a tharraingt as an téacs.

2 Mar go bhfuil léacsacan tuairime in úsáid san alt seo, úsáidtear focal tuairime agus focail tuairime san alt seo.

3 Níl an frása iomlán *sentiment lexicon* sa bhunachar téarmaíochta http://tearma.ie, ach tugtar *léacsacan* mar aistriúchán ar *lexicon*. Tugtar *dearcadh, tuairim* nó *barúil* don Ghaeilge ar *sentiment* san Fhoclóir Nua Béarla-Gaeilge. Úsáidtear *léacsacan tuairime* san alt seo.

4 Samplaí tógtha ón léacsacan ar líne, ag http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

ar fáil do mhórtheangacha an domhain, ach níl an freastal céanna déanta ar na mionteangacha. Maidir leis an nGaeilge, tá leagan Gaeilge den léacsacan AFINN ar fáil (Afli, 2017), ach níl leagan den léacsacan Bing aistrithe go Gaeilge. Mar sin, socraíodh, mar chuid den obair seo, an léacsacan Bing a aistriú go Gaeilge.

Léacsacan Tuairime Nua don Ghaeilge

Ag Cruthú Léacsacain Tuairime

Cruthaítear léacsacain tuairime ar bhealaí éagsúla, agus tá cur síos uileghabhálach den cheird in Liu (2012). Chun léacsacan tuairime cuimsitheach a chruthú, roghnaíonn an taighdeoir féin an liosta focal tuairime agus na luachanna a bhaineann leo, ach tá an próiseas seo an-fhada, agus tógann sé go leor acmhainní. Gan mórán acmhainní, is féidir léacsacain tuairime níos lú agus níos simplí a chruthú, agus feabhas a chur air ag úsáid modhanna uathoibríocha.

Tá cleasa éagsúla gur féidir a úsáid chun barr feabhais a chur ar léacsacan tuairime. Mar shampla, is féidir na haidiachtaí agus na cónaisc i mbunliosta focail tuairime a úsáid chun liosta níos cuimsithí a chruthú (Hatzivassiloglou & McKeown, 1997). Go minic i mBéarla, bíonn an tuairim chéanna ag focail atá ceangailte lena chéile leis an gcónasc *and*, agus go minic bíonn focail leis an malairt tuairime ceangailte lena chéile le *but*. Tugann Hatzivassiloglou & McKeown (1997) an sampla:

simple and well-received
simplistic but well-received

Is léir go bhfuil luach dearfach ag na focail *simple* agus *well-received* sa chéad abairt. Ach sa dara abairt tugann an focal *but* leid dúinn go bhfuil *simplistic* diúltach, ainneoin go bhfuil *well-received* dearfach san abairt seo freisin.

Dar ndóigh, nuair atá léacsacain tuairime cruthaithe do theanga amháin, is féidir ceann do theanga eile a chruthú tríd é a aistriú. Mar a luadh thuas, tá léacsacan Bing inaistrithe go Gaeilge, ach tá údar eile againn le Bing a roghnú don obair seo. Úsáidtear an pacáiste bogearraí *tidytext* (Silge and Robinson, 2016)[5] chun mianadóireacht tuairime a chur i bhfeidhm ar théacs Béarla ag úsáid léacsacain tuairime. Tá léacsacan Bing curtha in oiriúint do *tidytext*, rud a chabhraíonn go mór leis an bpróiseas anailíse.

5 Tá an pacáiste *tidytext* mar chuid den chreatlach ríomhchlárúcháin *R* (R Core Team, 2014), agus tá go leor buntáistí leis an gcreatlach sin. Mar shampla, is féidir giolcacha a bhailiú le pacáiste eile *R*, *twitteR* (Gentry, 2015).

Ag Aistriú Léacsacan Tuairime Bing go Gaeilge

Tá na mílte focal i léacsacain tuairime, agus tá iarrachtaí déanta teicníochtaí ríomhaistriúcháin a úsáid chun an t-ualach a laghdú ar an aistritheoir. Mar shampla, aistríodh léacsacan ó Bhéarla go hArabach ag úsáid ríomhaistriúcháin (Mohammed *et al*., 2016), ach dar leis na húdair, go mbíonn sé deacair tuairim na bhfocal a chosaint san aistriúchán. Mar sin, sheachain muid an ríomhaistriúchán agus thug muid faoi an 6788 téarma i léacsacan Bing a aistriú ar bhealach níos traidisiúnta.

Is obair aistriúcháin éagsúil é seo. Ní hamháin go bhfuil gach focal sa léacsacan le haistriú, ach chomh maith leis sin caithfear an luach tuairime céanna (*dearfach* nó *diúltach*) a tugadh don fhocal a chaomhnú san aistriúchán. Le gnáthaistriúchán téacs, faigheann an t-aistritheoir leid ó chomhthéacs na habairte, nó b'fhéidir ó réimse an téacs, a chuidíonn leis an aistriúchán. Ach san obair aistriúcháin seo, níl romhat ach an focal glan, gan chomhthéacs, seachas go bhfuil a fhios agat an focal *dearfach* nó *diúltach* é.

Mar shampla, tá an focal *adventurous* rangaithe mar fhocal *dearfach* i léacsacan Bing. San iontráil do *adventurous* ar Fhoclóir Nua Béarla-Gaeilge bíonn *fiontrach*, *dána*, agus *eachtrúil* le fáil mar aistriúcháin de. Sa chás seo, cuirtear *dána* as an iomaíocht, mar bíonn an focal sin diúltach an chuid is mó den am. Mar sin, gan leid ón gcomhthéacs, bhí sé i gcónaí mar sprioc an *sentiment* nó an tuairim a chosaint san aistriúchán.

Déileáiltear le focail chomhchiallacha ar bhealach ar leith. Is féidir an próiseas seo a léiriú ag úsáid an bhunfhocail Bhéarla *adorable*. Is féidir spriocfhocal leis an luach tuairime céanna, mar shampla *álainn*, a roghnú mar aistriúchán. Nuair a thiocfar ar bhunfhocal gaolmhar sa bhunléacsacan, mar shampla *beautiful*, is féidir foirm eile den fhocal *álainn* a roghnú mar aistriúchán, mar shampla *dóighiúil*.

Go hidéalach, d'aistreofaí gach foirm de na spriocfhocail agus gach téarma comhchiallach a bhaineann leo atá sa léacsacan Béarla. Ach níl an líon céanna téarmaí comhchiallacha i nGaeilge leis an luach tuairime céanna agus atá i mBéarla. Mar sin, is túisce go n-úsáidfí na téarmaí comhchiallacha Gaeilge ar fad agus go mbíonn focail Bhéarla ón léacsacan fágtha gan aistriúchán Gaeilge orthu dá thoradh seo.

Ní gá go mbeadh an t-aistriúchán an-dílis nuair atá leagan Gaeilge de Bing á chruthú. Níl muid ach ag úsáid an leagan Béarla de léacsacan Bing chun stór focal Gaeilge gaolmhar a ghiniúint do léacsacan Gaeilge Bing. Ní bheidh aon nasc, fisiciúil ná fíorúil, idir léacsacan Béarla Bing agus léacsacan Gaeilge Bing, ná idir na focail aonair sa dá léacsacan.

Ag deireadh an phróisis aistriúcháin, fágtar go raibh, sa chéad leagan den léacsacan aistrithe, 3744 focal tuairime Gaeilge, i gcomparáid le 6788 focal tuairime sa leagan Béarla. Tugtar, sa Tábla thíos, samplaí éagsúla de fhocail tuairime ó léacsacan tuairime Bing (Béarla) agus na focail tuairime comhfhreagracha sa léacsacan Gaeilge. I gcás *trustingly*, úsáideadh an focal *hiontaobhach* seachas *go hiontaobhach* mar go bhfuilimid ag plé le focal aonair.

Bing (Béarla)	Bing (Gaeilge)	Luach Tuairime
trusting	iontaobhach	dearfach
trustingly	hiontaobhach	dearfach
trivial	díspeag	diúltach
trivialize	díspeagadh	diúltach

Tábla 1: Samplaí éagsúla de focail tuairime ó léacsacan tuairime Bing (Béarla) agus na focail tuairime comhfhreagracha sa léacsacan Gaeilge.

An Léacsacan Gaeilge á chur i bhfeidhm ar Ghiolcacha

An Téacs Gaeilge

Tá pobal Gaeilge gníomhach ar an ardán Twitter, agus na mílte giolcach á seoladh gach bliain. Coinnítear taifead den stór focal seo ar Twitter, is féidir é a íoslódáil agus a ionramháil, agus faoi mar a thuigfeá, bíonn an téacs breac le tuairimí. Mar sin, shocraigh muid an léacsacan tuairime Gaeilge nua a chur i bhfeidhm ar an gcorpas seo. Cé go bhfuil cead tugtha giolcacha ó Twitter a ionramháil, caithfear cloí le téarmaí agus coinníollacha an chomhlachta sin, go háirithe mar a bhaineann sé le hábhar a fhoilsiú[6]. Pléadh sa litríocht na treoirlínte eiticiúla seo do thaighdeoirí (Williams *et al.*, 2017). Go bunúsach, is féidir téacs as giolcacha a fhoilsiú a fhad is gur as cuntais eagraíochtúla iad nó gur cuntais iad a bhaineann le pearsanra poiblí.

Don taighde seo, baineadh úsáid as an bpacáiste R *twitteR* (Gentry, 2015) chun giolcacha Gaeilge a bhailiú. Bailíodh giolcacha ar amlínte[7] @NuachtTG4 (nuacht teilifíse ó TG4), @tuairiscnuacht (Tuairisc.ie, nuachtán ar líne), @SportTG4 (Spórt ar TG4, nuacht spóirt) agus @NuachtRnG (nuacht raidió ó RTÉ Raidió na Gaeltachta) ar an 14 Eanáir, 2018. Roghnaíodh na cuntais Twitter seo mar shamplaí de chuntais nuachta atá scríofa i nGaeilge (den chuid is mó), ar scéalta polaitíochta, spóirt agus sóisialta.

Sa tacar sonraí a bailíodh, tá 12,603 giolc a foilsíodh idir 2010 go dtí 2017. Níl giolcacha ó gach amlíne ón tréimhse chéanna, mar shampla is ó 2017 agus ó 2018 a tháinig na giolcacha ar amlíne @tuairiscnuacht. Is ó chuntas @NuachtTG4 a tháinig giolcacha ó 2010. Maidir le struchtúr an tacair shonraí, tá iontráil do gach giolc le huimhir aitheantais, an t-am a cruthaíodh í, an téacs agus sonraí eile.

6 https://developer.twitter.com/en/docs/tweets/search/guides/build-standard-query

7 Is éard is amlíne sa chomhthéacs seo ná na giolcacha is déanaí ó chuntas Twitter ar leith.

Liosta focal coitianta

Sular féidir an léacsacan tuairime a chur i bhfeidhm ar an téacs ó Twitter, caithfear liosta focal coitianta (*stop words* i mBéarla) a ghiniúint. Is iad seo na focail choitianta atá le scagadh amach as an téacs sula ndéantar anailís air. Cuir i gcás, má tá taighdeoir ag iarraidh breathnú ar na focail nó ar na topaicí is tábhachtaí i sliocht, bunaithe ar mhinicíocht, beidh focail ar nós *tá, bhí, le* i gcónaí ar na focail is minice, mar is iad na focail is mó úsáide sa teanga. Mar a bhaineann sé le mianadóireacht téacs, níl mórán faisnéise sna focail seo. Má scagtar amach na focail is coitinne sa teanga ó théacs ar leith, beidh na topaicí faoi chaibidil níos feiceálaí sa liosta de na focail is minice, seachas na focail gan eolas tábhachtach.

Cruthaíodh liosta meaisín-inléite d'fhocail choitianta agus san áireamh sa liosta seo, tá: foirmeacha den bhriathar bí, réamhfhocail, forainmneacha, forainmneacha réamhfhoclacha[8]. Rinneadh próiseáil ar an téacs sna giolcacha sular úsáideadh an léacsacan tuairime: scagadh amach na focail choitianta sna giolcacha leis an liosta nua, agus scagadh amach freisin na hainmneacha de na cuntais Twitter, na haisclibeanna agus béarlagair in úsáid ar Twitter (dm, rt).

Ag forbairt uirlisí anailíse de théacs Gaeilge

Tá an teanga ríomhchlárúcháin *R* (R Core Team, 2014) go mór in úsáid san eolaíocht sonraí, agus is féidir roinnt de na tascanna a bhaineann le próiseáil téacs Gaeilge a dhéanamh leis an bpacáiste *R tidytext*. Is féidir próiseáil téacs Gaeilge a dhéanamh le creathlaigh ríomhchlárúcháin ar nós NLTK (Loper & Bird, 2002) nó GATE (Cunningham *et al*., 2002), ach san alt seo déantar an anailís ar fad le *tidytext* agus pacáistí éagsúla *R* mar gheall ar an oiriúnacht don obair seo. Le *R,* is féidir sonraí téacs a bhailiú, agus anailís a dhéanamh ar an téacs leis an gcreatlach ríomhchlárúcháin céanna go héasca.

Ionas gur féidir sonraí úsáideacha a roinnt le taighdeoirí eile, is féidir pacáiste bogearraí a fhorbairt le *R*[9], agus sonraí (in-atáirgthe, ar leagan amach néata) a chur ar fáil mar chuid dó. Ní mór don fhorbróir cur chuige sonrach a leanúint, mar shampla, is gá cáipéisí d'úsáideoirí a chur ar fáil agus bíonn formáid ar leith ag baint leo. Má leanann an forbróir an struchtúr sin, bíonn an pacáiste deiridh agus na sonraí iniata éasca le híoslódáil (ó sheirbhís ar nós GitHub[10] nó CRAN[11]) agus éasca le húsáid i bpáirt le pacáistí eile *R*.

Mar chuid den obair seo, críobhadh agus cuireadh ar fáil an pacáiste bogearraí *stad*, agus mar chuid dó, tá liosta focal coitianta i nGaeilge agus leagan den léacsacan aistrithe[12]. Baineadh úsáid as *stad* chun an anailís go léir, ina bhfuil cur síos thíos air, a dhéanamh.

8 Le fáil ag: https://github.com/cldatascience/stad.

9 http://r-pkgs.had.co.nz/package.html.

10 https://github.com/.

11 https://cran.r-project.org/.

12 Na focail tuairime ón léacsacan a bhí sa chorpas Gaeilge Twitter thíos (963 focal).

Anailís ar na Torthaí

Tuin na nGiolcach Aonair

Cuireadh an léacsacan nua sa phacáiste *stad* i bhfeidhm ar an téacs ó Twitter. Fuarthas scór tuairime do gach giolc, bunaithe ar na focail dhearfacha lúide na focail dhiúltacha mar chéatadán de na focail uilig sa ghiolc. Tugann sé seo meastachán dúinn ar thuin an ghiolc. Tugtar, sa Tábla thíos, samplaí d'abairtí simplí chun an scór sin a léiriú[13].

Abairt	Focail iomlán	Focail diúltacha	Focail neodracha	Focail dearfacha	Scór
Comhghairdeas, rinne tú éacht	3	0	1 (rinne)	2 (comhghairdeas, éacht)	66.66
Tá an stoirm ag teacht	2	1 (stoirm)	1 (teacht)	0	-50.00

Tábla 2: Samplaí d'abairtí a léiríonn meicníocht an scórála á fheidhmiú. Níl na focail choitianta san áireamh sa cholún "Focail Iomlán". Níl na focail neodracha sa léacsacan tuairime.

Dar leis an scór sin, tá na giolcacha seo níos diúltaí (giolc neamhathraithe):

> NuachtTG4 (NuachtTG4). "50,000 gan aibhléis le Stoirm Eleanor ag réabadh iarthar na tíre https://t.co/Zh2hupisPE https://t.co/uzkcoISDhg". 02 Jan 2018, 19:23 UTC. Tweet.

> Tuairisc.ie (tuairiscnuacht). "Tá roth an reifrinn ag casadh, agus achrann géar ag teannadh linn https://t.co/cxvl2ALrHF https://t.co/KVhCGzCPSn". 13 Jan 2018, 05:55 UTC. Tweet.

Agus tá na giolcacha seo níos dearfaí:

> TG4, Spórt (SportTG4). "Tá tallann den scoth sa tír seo agus chun sin a cheiliúradh beidh sraith nua ag tosú faoi mhná & cluichí CLG. #TG4 https://t.co/2Lf8dKrEIm". 09 Jan 2017, 20:12 UTC. Tweet.

> TG4, Spórt (SportTG4). "Beidh na buaicphointí againn Dé Sathairn @Munsterrugby @connachtrugby #Irishrugby https://t.co/GCCpZPnVzc". 07 Dec 2017, 16:00 UTC. Tweet.

13 Cód samplach san fhillteán *tests* sa phacáiste *stad.*

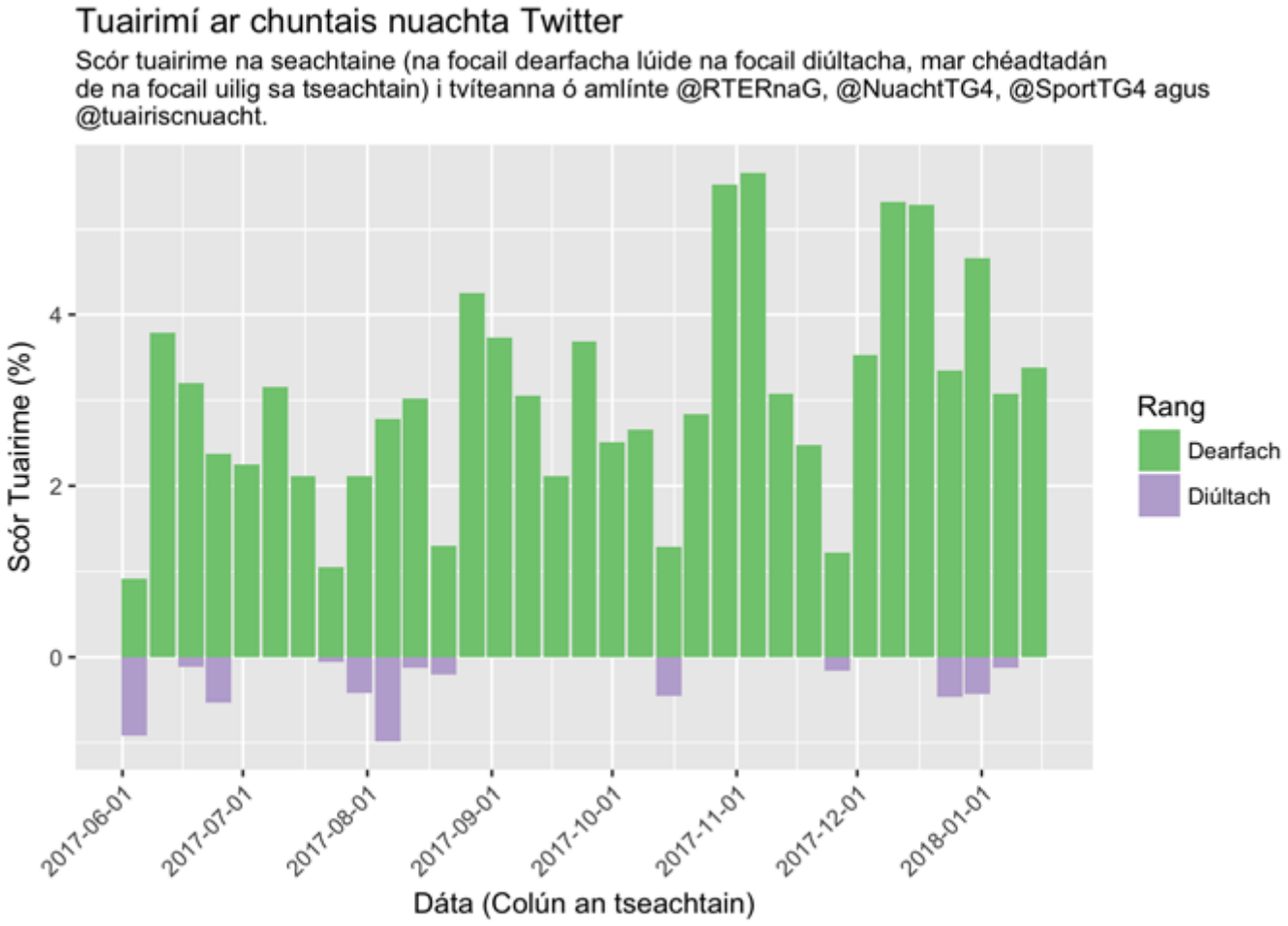
> Tuairisc.ie (tuairiscnuacht). "Mol an óige agus tiocfaidh sí, ach an bhfanfaidh sí? https://t.co/9x6SwzmdDW". 26 Aug 2017, 16:00 UTC. Tweet.

Is cinnte go bhfuil an chéad ghiolc ó @SportTG4 dearfach, ach sa dara giolc bíonn brí sách neodrach leis an bhfocal *buaicphointí* i comhthéacs an spóirt. Maidir leis an ngiolc ó @tuairiscnuacht, tá na focail *mol* agus *óige* dearfach sa léacsacan nua Gaeilge ach is léir go bhfuil tionchar láidir ag an gceist atá curtha ag deireadh na habairte ar thuin na giolcaí.

Is í seo an chéad leagan den léacsacan Gaeilge, agus léiríonn an giolc deiridh seo na srianta atá leis. Mar shampla, ní thógtar san áireamh, agus an scór á ríomh, an comhartha ceiste, ná an focal *ach* san abairt. Is léir go bhfuil tionchar acu seo ar thuin iomlán na giolcaí.

Tuairimí na seachtaine

Chomh maith le scór tuairime do gach giolc, rinneadh grúpaí de na giolcacha a cruthaíodh gach seachtain agus comhaireadh líon na bhfocal dearfach agus líon na bhfocal diúltach gach seachtain. Fuarthas scór tuairime na seachtaine, bunaithe ar na focail dhearfacha lúide na focail dhiúltacha, mar chéatadán de na focail uilig sa tseachtain (féach Léaráid 1). Feictear sa léaráid go bhfuil tuairim dhearfach den chuid is mó ag téacs na gcuntas Twitter sin. Léiríonn sé seo go bhfuil níos mo téarmaí dearfacha ón léacsacan in úsáid i gcomparáid le téarmaí diúltacha ar na cuntais Twitter seo. Ní haon iontas é seo, mar gur cuntais chorparáideacha iad, agus baineann siad úsáid astu chun poiblíocht a dhéanamh ar a gcuid clár. Tá patrúin shéasúracha le feiceáil freisin. Feictear, mar shampla, ardú sa scór tuairime le linn Oireachtas na Gaeilge.



Léaráid 1: Scór tuairime seachtainiúil ar chuntais nuachta Twitter.

Obair don Todhchaí

An léacsacan a mheas go foirmeálta

Is próiseas atriallach atá i bhforbairt léacsacan tuairime, agus níl sa leagan Gaeilge seo de léacsacan Bing ach an chéad timthriall. Is féidir cur leis an léacsacan agus é a fheabhsú, ach caithfear gach leagan den léacsacan tuairime a mheas go cainníochtúil le turgnamh foirmeálta. Le bheith úsáideach mar uirlis chainníochtúil, caithfear cruinneas agus críochnúlacht an léacsacain a thomhais. Is é sin le rá, an bhfuil luach tuairime na bhfocal ceart sa léacsacan aistrithe? An bhfuil na focail is tábhachtaí ó thaobh tuairime de sa léacsacan? An bhfuil an léacsacan ionadaíoch ar an téarmaíocht Ghaeilge atá in úsáid ar Twitter?

Chun an obair sin a chur i gcrích, caithfear téacs Gaeilge a bhailiú agus meastóirí a fháil chun luach tuairime a thabhairt do gach aonad sa téacs (i.e. gach téacs, gach léirmheas, agus araile). Tá dhá rud le sárú anseo, meastóirí a earcú agus maoiniú a aimsiú le hobair na meastóirí a íoc, murar féidir an obair a dhéanamh

go deonach. Go minic le turgnamh ar théacs Béarla, úsáidtear seirbhís ar nós *CrowdFlower*[14] agus íoctar as meastóirí chun an aicmiú a dhéanamh. Bíonn meithleacha ar fáil do chuid de na mórtheangacha (Béarla, Spáinnis, Hiondúis, Fraincis agus araile) ach ní féidir meithleacha Gaeilge a fháil ón tseirbhís dhíreach sin. Pléadh obair mheithle i comhthéacs mionteangacha in alt Dowling. (2017), agus tugadh cur síos ar thomhas léacsacan tuairime Gaeilge in Afli *et al.* (2017).

Ag cur breis focal leis an Léacsacan

San alt seo, rinneadh léacsacan tuairime a aistriú go Gaeilge, ach is féidir cur leis ar bhealach níos struchtúrtha freisin, agus na hearráidí a bhaint amach chomh maith. Chonaic muid go raibh raon leathan focal in úsáid ar Twitter nach raibh le fáil sa léacsacan. I measc na bhfocal seo bhí na focail a thosaíonn le *sár, rí* nó *ard*, (focail dhearfacha) agus focail a thosaíonn le *frith, droch, neamh* nó *mí* (focail dhiúltacha). D'fhéadfaí iad seo a áireamh sa chéad leagan eile den léacsacan tuairime. Freisin, níl gach sampla de na focail dhiúltacha sin sa bhunléacsacan. Cuir i gcás *frithdhaonlathachas* nó *drochstaid,* focail nach raibh sa léacsacan, ach a bhí sa chorpas:

> Tuairisc.ie (tuairiscnuacht). "Más cúis imní an **frithdhaonlathachas** i gcás na Catalóine, is amhlaidh i gcás Thír na mBascach… https://t.co/mTgmpcsaDh". 31 Oct 2017, 06:46 UTC. Tweet

> NuachtTG4 (NuachtTG4). "Cartlann @TG4TV: **Drochstaid** na scoile ar Oileán Árainn Mhór - i dTír Chonaill a bhí faoi chaibidil ar an Nuacht. https://t.co/5cQ1KenxSy". 28 Nov 2016, 20:11 UTC. Tweet

14 https://www.crowdflower.com/

Conclúid

San alt seo rinneadh cur síos ar an gcaoi ar féidir léacsacan tuairime a úsáid chun anailís a dhéanamh ar dhearcthaí an phobail atá gníomhach ar na meáin shóisialta Gaeilge. Míníodh an chaoi ar cruthaíodh léacsacan tuairime nua don Ghaeilge. Mar chuid den obair seo cruthaíodh pacáiste bogearra nua, *stad*, chun gur féidir an léacsacan a chur i bhfeidhm ar chorpas téacs agus anailís a dhéanamh air ag úsáid modhanna de chuid na mianadóireachta tuairime. Cruthaíodh agus cuireadh foinsí meaisín-inléite ar fáil, mar shampla liosta d'fhocail choitianta i nGaeilge agus leagan aistriúcháin Ghaeilge den léacsacan tuairime Bing. Léiríodh na deiseanna a thugann Twitter do thaighdeoirí agus iad ag iarraidh téacs feiliúnach a bhailiú. Ar an ábhar sin, bailíodh giolcacha ó na meáin nuachta ar Twitter agus rinneadh anailís ar na tuairimí sna giolcacha leis an léacsacan.

Léiríonn an triail seo go bhfuil an anailís seo indéanta ar théacs Gaeilge le huirlisí atá ar fáil go hoscailte. Mar sin féin, tá dúshláin le sárú chomh maith. Le léacsacan tuairime a mheas go cainníochtúil, bíonn turgnamh foirmeálta riachtanach. Leis an turgnamh sin a dhéanamh, bíonn téacs Gaeilge ag teastáil, chomh maith le meastóirí chun luach tuairime a thabhairt do gach uile sliocht den téacs.

Mar fhocal scoir, tá na foinsí agus na huirlisí a cruthaíodh don alt seo ar fáil go hoscailte. Ina measc, tá leagan den léacsacan Gaeilge Bing, an pacáiste anailíse *stad*, agus liosta de na focail choitianta Gaeilge.

Leabharliosta

Leabhair

Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

Ailt

Afli, H., McGuire, S. & Way, A., (2017) 'Sentiment Translation for low resourced languages: Experiments on Irish General Election Tweets.' *The 18th International Conference on Intelligent Text Processing and Computational Linguistics*. Budapest, Hungary.

Arcan, M. *et al.*, (2016) 'IRIS: English-Irish Machine Translation System.' *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2016). Portorož, Slovenia.

Cunningham, H. *et al.*, (2002) 'GATE: An Architecture for Development of Robust HLT Applications.' *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL '02). 168-75.

Dowling, M. *et al.*, (2015) 'Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish', in *The 4th LRL Workshop: 'Language Technologies in support of Less-Resourced Languages'*. Poznan, Poland.

Dowling, M., Lynn, T. & Way, A., (2017) 'A Crowd-sourcing Approach for Translations of Minority Language User-Generated Content (UGC).' *Proceedings of 1st Workshop on Social MT*. Prague, Czech Republic.

Hatzivassiloglou, V. & McKeown, K.R., (1997) 'Predicting the semantic orientation of adjectives.' *Proceedings of the 35th annual meeting on Association for Computational Linguistics*.

Hu, M. & Liu, B. (2004) 'Mining and summarizing customer reviews', in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*. 168-77.

Judge, J. Ní Chasaide, A., Ní Dhubhda, R., Scannell, K.P., & Uí Dhonnchadha, E., (2012) 'The Irish Language in the Digital Age.' Rehm, G. & Uszkoreit, H. (eag.). Springer (META-NET White Paper Series: Europe's Languages in the Digital Age).

Loper, E. & Bird, S., (2002) 'NLTK: The Natural Language Toolkit.' *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics – Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics (ETMTNLP '02) 63-70.

Lynn, T., Foster, J. & Dras, M., (2017) 'Morphological Features of the Irish Universal Dependency Treebank.' *15th International Workshop on Treebanks and Linguistic Theories*. Bloomington, Indiana.

Mohammad, S.M. & Turney, P.D., (2010) 'Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon.' *CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. 26-34.

Mohammad, S.M., Salameh, M. & Kiritchenko, S., (2016) 'How translation alters sentiment', *Journal of Artificial Intelligence Research*, 55. 95-130.

Nielsen, F.Å., (2011) 'A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.' *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*. 93-8.

Silge, J. & Robinson, D. (2016) 'tidytext: Text Mining and Analysis Using Tidy Data Principles in R', *JOSS*. The Open Journal, 1(3).

Williams, M.L., Burnap, P., & Sloan, L., (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*.

Foinsí Leictreonacha

Gentry, J., (2015) 'twitteR: R Based Twitter Client'. Le fáil ag: https://cran.r-project.org/package=twitteR. (Léite: 29 Aibreán 2018).

R Core Team, (2014) 'R: A Language and Environment for Statistical Computing', *R Foundation for Statistical Computing*, Vienna, Austria. Le fáil ag: http://www.R-project.org/. (Léite: 29 Aibreán 2018).

Wickham, H. (2017) 'tidyverse: Easily Install and Load the "Tidyverse"'. Le fáil ag: https://cran.r-project.org/package=tidyverse. (Léite: 29 Aibreán 2018).