Txiolari euskaldunen hizkuntza-erabilera

Definizioa

CodeSyntaxek 2010ean sortu zuen UMAP tresna eta 2016an berregituratu eta oraingo egiturara moldatu zen. Modu automatiko-robotikoan Twitterreko euskarazko jarduna jasotzen du. Algoritmo baten bitartez Twitterreko euskaldun potentzialak detektatzen ditu, eta horien artean euskaraz egiten dutenen jarraipena egiten du. Erabiltzaile euskaldun eta aktibo horien txio guztiak analizatzen dira etengabe, eta txio bakoitzaren hizkuntza detektatu eta gordetzen da.

Hedadura

Aztertzen den lagina Twitterren euskaraz egiten duten erabiltzaileena da. Umapeko datu-basean 60.000 euskaldun potentzial daude identifikatuta (2020ko urtarrilean), baina horietako gutxi batzuk bakarrik dira Twitterren aktiboak eta euskaraz egin dutenak. 10 urte hauetan 22.000 euskal txiolari detektatu ditugu (euskaraz egin dutenak). Gainontzekoak inaktiboak dira, edo kontu pribatua dute (aztertu ezinekoa), edo ez dute sekula euskaraz egin (gehienek euskara, gutxienez, ulertu egingo duten arren).
Jaso eta ikertu diren txioak, beraz, azken 10 urteotan inoiz euskaraz egin duten txiolarienak dira. Jakin badakigu detektatutako 22.000 erabiltzaile inguru horiek ez direla izango ez txiolari euskaldun guztiak, ezta inoiz txioren bat euskaraz argitaratu duten guztiak ere. Hala ere, euskaraz maiztasun minimo batekin txiokatzen duten erabiltzaile gehien-gehienak identifikatzen direla ziur gaude. Baliteke %5-10 inguruko errorea egotea, asko jota, baina Twitterreko euskal komunitatearen hizkuntza-erabileren argazki zehatza lortzen dugu, eta urtez urteko jarraipena egiteko datu-serie sendoa.

Aldagaiak eta kategoriak

Hizkuntza-erabileraren jarraipena (euskara, gaztelania, frantsesa, ingelesa eta sailkatu gabekoak*). Txio kopurua eta hizkuntza bakoitzaren erabilera (%).

Euskara-erabilera sexuaren arabera: Hizkuntza-erabileraren jarraipena (euskara, gaztelania, frantsesa eta ingelesa) generoaren arabera, emakume eta gizonezkoen arteko ezberdintasuna aztertzeko.

* Txioen hizkuntza automatikoki detektatzen da. Ez du positibo faltsurik ematen. 4 hizkuntza nagusi identifikatzen ditugu. “Sailkatu gabe” kategorian sartzen dira: 1) beste hizkuntza batzuk (katalana, italiera, alemaniera...); 2) mezu elebidunak (euskara-erdara nahastean); 3) mezu hiperlaburrak edo testurik gabeak (emotikokonoak, gif animatuak...); 4) bestelakoak.

Azpi-adierazleak (gurutzaketak)	
Iturria	
Umap.eus - Codesyntax	
Iturriaren ezaugarriak	Urtero kalkulatzen dira datuak abenduaren 31an jasotako txioekin
Iturriaren arduraduna	CodeSyntax S.L.
Lehenengo datua eta maiztasuna	• Tresna 2010ean garatu zen. • Harrezkero, erabiltzaileen detekzioa eta txioen parseoa (analisia) etengabe egin da, minuturo minuturo krona exekutatuz. • Gainera, erabiltzaileen parseoa atzerantz ere egiten denez, erabiltzaile batzuen kasuan 2007tik honako datu guztiak ditugu. • Guztira 40 milioi txio jaso eta aztertu dira (2020ko urtarrileko datuak). • Urtekako datuei dagokionez, serie historikoa sendoa da 2014az geroztik eta zehatzagoa 2016az geroztik (erreminta berregokitu baitzen 2016an).
Unibertsoa	2018an 20.191 txiolari, 2019an 21.587 txiolari eta 2020an 24.161 txiolari
Erabiltzaile aktiboak	Erabiltzaile aktiboak urtean zehar txioren bat egin dutenak dira: 2018an 14.240, 2019an 14.354 eta 2020an 15.992
Lagina eta akats-tartea	60.000 Twitter kontu aztertu dira. Horietatik 22.000 inguru izan dira inoiz euskaraz aktiboki txiokatu dutenak eta kontu horien txioak dira jaso eta aztertu direnak.
Informazio osagarria	
Argitalpenak eta txostenak	Umap.eus tresnari buruz: https://umap.eus/faq/ Hizkuntza detekzioari buruz: https://umap.eus/albisteak/umapeko-hizkuntza-detektoreaz