



IKER
GAZTE
NAZIOARTEKO
IKERKETA EUSKARAZ

III. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2019ko maiatzaren 27, 28 eta 29
Baiona, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZARTE ZIENTZIAK ETA ZUZENBIDEA

Euskaldun gazte eta helduen
harremanak Twitterren

*Joseba Fernandez de Landa,
Rodrigo Agerri eta Iñaki Alegria*

83-90 or.

<https://dx.doi.org/10.26876/ikergazte.iii.02.11>



Euskaldun gazte eta helduen harremanak Twitterren

Fernandez de Landa, Joseba; Agerri, Rodrigo eta Alegria, Iñaki

IXA Taldea, Euskal Herriko Unibertsitatea (UPV/EHU), Donostia-San Sebastián

joseba.fdl@gmail.com

Laburpena

Sare sozialek mugarik gabeko komunikazioa ahalbidetzeaz gain, ikerketa sozial eta linguistikorako tresna ere badira, datu kopuru haundiak lortzeko iturri garrantzitsu bat direlako. Lan honen asmoa txiolari euskaldun gazte eta helduen harremanak aztertzea izango da, horretarako konputazio zientzietako teknikak ikerketa sozialean aplikatuz. Horretarako, 8000 erabiltzailearen euskarazko 6 milioi txio baino gehiago lortu dira. Lehenengo, erabiltzaile bakoitza adinaren (gazte/heldu) arabera sailkatuko dugu, horretarako txio bakoitzaren idazkera estiloa formala edo informala den zehaztuz eta ikasketa automatikoa erabiliz. Bigarren pausuan, birtxioak erabiliz, erabiltzaileen arteko harremanak zeintzuk diren azaleratu eta azpi-komunitateak nola osatzen dituzten erakutsiko da.

Hitz gakoak: sare sozialak, gazteak, harremanak, euskara

Abstract

Social networks allow communication without limit, but they are also a useful tool for social and linguistic research, allowing us to collect big amounts of unstructured data. In this work, the aim will be to study the social relations of young and adult basque speakers' relations based on the tweets they generate, applying computer science techniques to social research. With this aim in mind, we have gathered over 6 million tweets written in Basque from near 8000 users. First, we classified each user's tweet by age (young/adult) by establishing whether the writing style of each tweet is formal/informal, using Machine Learning methods. Second, we establish the relations that emerge between the users based on their retweets, characterizing also the various communities that arise from the analysis of the data.

Keywords: social networks, young people, relations, basque language

1. Sarrera eta motibazioa

Sare sozialetan euskal hiztun gazteek dauzkaten harreman sareen edo komunitateen analisisa egitea izan dira lan honen helburua. Sare sozialak gazteon esparru bezala kontsideratu ditzakegu, haien sozializaziorako lanabes garrantzitsu bat izanik. Modu honetan, sareetan egiten den euskararen erabilera ezagutzea oso interesgarria da, garai berrietara moldatzeko euskarak daukan ahalmena ikusi eta belaunaldi berrien espazioa diren sare sozialetan nola gauzatzen den ezagutzeko asmoz. Honela, euskararen egoera ezagutzetik gertuago egoteko aukera edukiko da, ikerketa-teknika tradizionalen osagarria izango den begirada berria eskainiz. Helburu horretarako Hizkuntza-ren prozesamenduko tekniketan oinarritu gara, informatikako teknologiak gizarte-ikerkuntzan aplikatuz. Gizarte-zientzien eta konputazio-zientzien arteko konbinazioan kokatzen da lan honen ekarpen nagusia, aurrerapen teknologikoetan eta informazio mugagabearen oinarritutako egungo gizartea interpretatzen eta ulertzen lagunduko duen sinbiosia.

Honela, Twitter sare sozialean euskal komunitatearen inguruko ikerketa burutu da, sare sozialak informazio iturri garrantzitsu bat direla frogatuz. Era honetan, euskal txiolarien komunitatearen euskarazko 6 milioi txio baino gehiago lortu dira, ia 8000 erabiltzaile ezberdinetatik. Datu-base erraldoi honetatik (Big Data), interpretagarria den informazioa erauztea izango da asmoa, desegituratutako datu hauek ulergarri bihurtuz (Data Mining). Behin datuak erauzi direla, hurrengo pausua gazteak detektatzean oinarrituko da, adina bezalako ezaugarri demografikoak iradokitze saiakerak eginez, horretarako lengoia naturalaren prozesamenduko (NLP) teknikak erabiliz. Honela, txio pertsonalen idazteko eran oinarrituz, erabiltzaile gazteak eta helduak ezberdindu dira hizkuntza ez-formalaren

erabileran oinarrituz. Erabiltzaileak gazte eta heldu artean ezberdinduta daudenean, hauek norekin harremantzen diren argitzea izango da azken pausua. Harreman hauek nola ematen diren argitzeko, birtxioetan oinarrituta, komunitateak zehatz eta hauetako pertsona garrantzitsuak identifikatuko dira komunitate hauek definitzeko. Era honetan, helduen zein gazteen taldeen barne komunitateak argitara eman eta konparatu ahal izango dira eta sare sozialetan euskaraz nola harremantzen garen ezagutu.

2. Arloko egoera eta ikerketaren helburuak

Sare sozialen arloan ezaugarri demografikoak zeintzuk diren iragartzea, *Social Media Mining* arloan gai ikertuetako bat da. Era honetan, sare sozialetako erabiltzaileen ezaugarriak antzemateko hainbat sistema garatu dira, adina edota generoa izanik ohikoenak (Cesare et al., 2017). Soziolinguistikaren aburuz, Ahoz edo idatziz adierazteko estiloak komunitate zehatz batekiko pertenezia adierazi dezake, besteak beste, adin talde batekiko pertenezia (Nguyen et al., 2016). Kasu zehatz honetan, Twitterreko erabiltzaile euskaldunen bizitza etapa identifikatzea izango da asmoa, erabiltzaile hauek gazte edo helduen artean ezberdindzea, horretarako testua baliatzea posiblea dela erakutsiz.

Bestalde, hainbat dira Twitter sare sozialeko birtxioetan oinarrituta, komunitateak antzeman dituzten ikerketak. Horien artean nabarmendu genitzake polarizazio politikoa aztertzeke erabilitakoa Conover et al. (2011) eta baita independentziaren aldeko mugimenduetan egindakoa (Zubiaga et al., 2017). Ikerketa hauetan erakusten da, erabiltzaileek egindako birtxioetan oinarrituta, badagoela komunitateak edo taldeak iragartzeko aukera. Beraz, lan zehatz honetan, metodologia antzekoak erabiliko ditugu, euskal komunitatearen baitan komunitateak nola ematen diren aurreikusteko, hau da, harremanak nola ematen diren ikusteko.

Aukerak behin aztertuta, lan honen helburu orokorra, txiolari gazte eta heldu euskaldunen harremanak nola ematen diren konparatu eta aztertzea izango da. Era honetan, gazteen errealitate ezezagunera hurbilpen bat lortzeaz gain, XXI. mendeko erronketara euskara nola egokitzen ari den ezagutu ahal da. Horretarako, lehenengo azpi-helburua txiolari euskaldunak identifikatzea izango da eta beren euskarazko txio pertsonal zein birtxioak jaso. Bigarren azpi-helburua, lortutako erabiltzaile euskaldunetan gazteak eta helduak identifikatzea izango da, idazkera estiloan oinarrituz, ezaugarri demografikoak iradokitzeke hurbilpena eginez. Hirugarren eta azken azpi-helburua, gazteen eta helduen harremanak zeintzuk diren identifikatu eta konparatzea izango da, euskal txiolarien eduki trukea baliatuz. Bestalde, esan beharra dago ere, lan honen bigarren mailako helburu bezala, gizarte zientzien eta konputazio zientzien arteko harremana sendotzea dela, gizarte ikerketarako bien arteko sinbiosia bilatuz.

3. Ikerketaren muina

Euskal txiolari gazte eta helduen errealitatea aztertuko duen ikerketa lan hau burutzeko, lana hiru zati ezberdinetan zatitzea erabaki da. Lehenik eta behin datuen erauzketa gauzatu beharko da, Twitter sare sozialean euskal erabiltzaileak identifikatu eta hauen datuak batu. Bigarren pausua, gazteak eta helduak bereiztean datza, bizitza etapen arteko konparaketa egitea interesatzen baitzaigu. Hirugarrenez eta azkenik, euskal txiolarien harremanak ezagutzeko, euskarazko birtxioetan oinarrituko gara, hauen atzean dagoen harremanen-sarea islatuz.

3.1. Datuen erauzketa

Komunitate zehatz baten azterketa egitera bidean beharrezkoa da komunitate hau identifikatzea, lehenik eta behin aztertu beharreko unibertsoa zehaztuz. Hau da, ikertu behar diren subjektuak identifikatzea, kasu honetan unibertsoa euskaraz egiten duten txiolariak izanik. Honela, Twitterreko sare erraldoian euskarazko erabiltzaileak topatzeari ekin zaio, horretarako, euskaraz txio kopuru minimo bat argitaratzen duten erabiltzaileen identifikazioa burutuz. Horrela, euskarazko txioak argitaratu dituzten subjektuetatik, euskal erabiltzaileak hautemateko, euskarazko txio kopuru minimo bat duten erabiltzaileak hautatu dira. Euskarazko twitter erabiltzaileak lortzeko, umap.eus webguneko euskal txiolarien zerrenda erabili da. Webgune honek garatutako sistema bati esker lortzen da euskal txiolarien zerrenda, euskal txiolariak detektatzeko gaitasuna baitu sistema horrek. Umap.eus webguneko sistema horrek gutxienez txioen %20a euskaraz argitaratzen duten erabiltzaileak barneratzen ditu zerrendan (Umap.eus, 2018). Era honetan, euskal txiolarien zerrenda lortu da, 8.189 erabiltzaile euskaldunek osatzen dutena.

Unibertsoa definituta dagoelarik, bigarren pausoa datuen erauzketa nola burutu definitzea izango da. Datuen erauzketa burutu ahal izateko Twitterreko APIa erabili da, honetarako Pythoneko *tweepy* (Roesslein, 2009) paketea hautatuz. Datu-bilketarako, erabiltzaileen erauzketa erabili da, euskal erabiltzaileen zerrenda baitaukagu. Honek esan nahi du zerrendako erabiltzaile bakoitzaren txioak erauziak izango direla, erabiltzaile bakoitzetik azkeneko

3.200 txioak lortuz gehienez jota. Erauzketa teknika hau aplikatuz, datuen bilketa burutu da, 2018ko maiatzaren 30 eta 31an. Orotara 7.980 euskal txiolarien azkeneko txioak erauztea lortu da, guztira 10 milioi txio baino gehiago batuz, hauetatik 6 milioi euskaraz.

1. taula. Erauzitako datu erabilgarriak, euskarazko txioen nolakotasuna.

	Euskarazko txio pertsonalak	Euskarazko birtxioak
Txio kopurua	3.171.785	2.891.136
Unitate lexikoak (terminoak)	1.434.050	813.833
Token kopurua (hitzak)	37.350.268	39.329.204

3.2. Gazte eta heldu artean bereizi

Esan bezala, erauzitako erabiltzaileak gazteak edo helduak diren antzematea izango da lanaren hurrengo pausoa, gerora bizitza etapa ezberdin hauetan harremanak nola ematen diren konparatzeko asmoarekin. Soziolinguistika habiapuntutzat hartuz (Nguyen *et al.*, 2016), ezaugarri demografikoak iradokitzeke saiakera egingo da, horretarako txioen testua kontutan hartuz. Erabiltzaile baten adina iradokitzeke ezaugarri garrantzitsuen adierazgarria idazkeran oinarritzen da (Rao *et al.*, 2010; Al Zamal *et al.*, 2012; Nguyen *et al.*, 2013; Morgan-Lopez *et al.*, 2017) eta horretan zentratzea erabaki da lan honetan. Honela, helduek gazteak baino hitz konbentzionalagoak erabiltzen dituzte (Nguyen *et al.*, 2013), hizkien errepikapena nabarmen gehiago ematen da erabiltzaile gazteen artean (Rao *et al.*, 2010; Rosenthal eta McKeown, 2011) eta hiztegiz kanpoko hitzak ohikoagoak dira gazteen artean (Rosenthal eta McKeown, 2011; Morgan-Lopez *et al.*, 2017). Ondorio orokorragoetara joz, idazkera aldatu egiten da adinean aurrera egin ahala, gazteagoek estilo ezohikoagoa edo informala daukatelarik (Nguyen *et al.*, 2013). Era honetan, gazteen idazteko modua, idazkera formal batetik gehien aldentzen dena bezala kontsideratuko da, helduen idazkera estilo formalarekin erlazionatuz eta gazteena estilo informalarekin. Horrela, euskal erabiltzaileak etiketatatu ordez, txioak etiketatzeari ekingo zaio, formal eta informal artean desberdinduz. Era honetan, txioen motaren kontzentrazioaren arabera sailkatuko dira erabiltzaileak gazte eta heldu artean.

Hortaz, gazte eta heldu klaseak iragartzeko, idatzitako testuaren estiloa izango dugu oinarri, txioen estiloa zehazki esanda. Era honetan, testuaren estiloaren arabera bi testu mota ezberdinduko dira, formal eta informala. Sailketaile beraz, txio informal zein formalak sailkatzean oinarrituko da, erabiltzaile bakoitzaren euskarazko txio guztiak honen arabera sailkatuz. Horrela, euskarazko txio pertsonal bakoitzaren testuan oinarrituko da gure sailketaileak, testua nola idatzia izan den aztertuz, formal edo informal moduan etiketatuko duena. Erabiltzaileak testu motaren arabera sailkatuko direnez, euskarazko txio formal edo informalaren kopuruaren arabera determinatuko da gazteak edo helduak diren. Honela, erabiltzaile baten euskarazko txioen gehiengo informala denean, erabiltzaileak gaztea dela kontsideratuko da, euskarazko txioen gehiengo formal denean oster, erabiltzaile heldu bezala sailkatuko da. Era honetan, erabiltzaileak ez dira itxura fisiko edo adinaren arabera sailkatuko, idazteko eragatik antzeman daitekeen izaeraren arabera baizik.

Euskarazko txioak formal eta informal artean sailkatzeko, ikasketa automatikoa oinarritutako *IXA pipes*en dokumentu sailketaileak (Agerri eta Rigau, 2016) erabiliko da. Metodo hau, ikasketa automatikoa oinarritua dago, baina beste datu-iturri batzuk lagungarri ditu ere. Sistema honek informazio lokala konbinatzen du etiketatutako gabe-ko testu kantitate handietatik indutitutako eta semantika distribuzionalean oinarritutako ezaugarrien clusterrekin (Agerri eta Rigau, 2016). Metodo honi esker, eskuz etiketatutako beharreko datu kopuru handiekiko dependentzia arindu egiten da, etiketatutako corpus txikiekin ere sailkapen egoki bat egitea ahalbidetuz. Honela, sistemaren garapen eta ebaluaketarako, sistemak ikas dezan eta ikasitakoa ondo aplikatzen duen ikusteko, 1.000 txio eskuz etiketatutako dira. Euskarazko 1.000 txio hauek aukeratu dira eta txio batek *formal* etiketa eramango du, hizkuntza estandarrean idatzia izan bada, edo *informal* etiketa, txioa modu kolokialean idatzia izan denean. Etiketatutako corpusa orekatua da, hau da, erdia txio informalez osatuta dago eta beste erdia txio formalez osatuta dago.

Sistema aukeratuta eta etiketatutako corpusa prest dagoenean, sistema garatu eta ebaluatzeari ekingo zaio. Honela, garapen fasean, etiketatutako 1.000 txioak baliatuz, *IXA pipes*en dokumentu sailketailean ezaugarri egokienak aukeratu dira eta errore gutxien ematen dituzten ezaugarriak aukeratu dira, horretarako 5 *fold*etako *cross validation* erabilita, etiketatutako corpusaren tamaina txikia izanda gainsorkuntza ekiditeko. Behin garapen fasea amaituta dagoela, aukeratutako ezaugarrien arabera eredu bat entrenatu eta ebaluatu da. Ebaluazioa burutzeko, aukeratutako ezaugarriak oinarri hartuta, eredu bat entrenatutako dugu horretarako etiketatutako 650 txio erabiliz. Behin ereduaren entrenamendua dagoela, ebaluazioari ekingo zaio, horretarako etiketatutako corpusetik entrenamendurako hartu ez diren 350 txioekin testa burutuz. Era honetan, sailketailearen fidagarritasuna frogatu ahal

da, esleitutako atazaren egokitasuna azaleraziz. Testa edo ebaluaketa burututa, ikusi da Asmatze tasa % 86,57koa dela, hau da, sistemak % 86,57ko aukera duela txio bat ondo sailkatzeko informal edo formal artean¹.

Sortutako sistema, ariketa praktiko erreal batean aplikatu da, txiolari gazteak eta helduak desberdintzeko erabiliko dena idazteko eran oinarrituz. Honela, erabiltzaile bakoitzaren txioak banatuko dira formal eta informal artean. Erabiltzaileen txio informalaren kontzentrazioan oinarrituta, gazte eta heldu artean ezberdindu dira, adinaren arabera egin ordez idazteko moduaren arabera eginez. Honela, txio informalaren kontzentrazioan oinarrituko dugu erabiltzaile baten gaztetasuna. Hasiera batean txioen % 50a informala izatea ezarri zen erabiltzailea gaztetzat jotzeko. Hala ere, azterketa kualitatibo sinple bat eginda, txio informalaren muga % 45era jaistea erabaki zen, horrela erabiltzaile gazteen kopuru minimo bat bermatzen delarik. Erabiltzailea gaztea dela kontsideratuko da txioen % 45a baino gehiago txio informalena baldin bada. Bestalde, txioen % 45a baino gutxiago informala baldin bada, erabiltzaile heldutzat joko da erabiltzaile zehatz hori. Honela, gure metodologia aplikatu da txio informalaren bitartez erabiltzaile gazteak antzemanen, adina igartzeko zailtasuna alboratuz eta lasterbide metodologikoa aplikatuz. Era honetan, erauzitako 7.980 euskal txiolarietatik, 7.087 erabiltzaile sailkatzea lortu da, hauetatik 5.508 erabiltzaile heldu bezala eta 1.579 gazte moduan kontsideratuz. Helduei dagokienez, 8 milioi txioz osatutako corpusa osatu da, hauetatik birtxioak 4.345.500 (% 56 euskaraz) eta txio pertsonalak 4.046.512 (% 65 euskaraz) izanik. Bestalde, gazteen corpusa 2 milioi txioz osatuta dago, hauetatik birtxioak 963.668 (% 42 euskaraz) eta txio pertsonalak 1.128.124 (% 47 euskaraz) izanik. Behin bizitza etaparen (gazte/heldu) eta hizkuntzaren arabera banatu direla txioak, euskal txiolarien harremanak nola ematen diren plazaratuko da, horretarako euskarazko birtxioak erabiliko direlarik.

2. taula. Erauzitako datuak bizitza etaparen arabera.

	Helduak	Gazteak
Erabiltzaileak	5.508	1.579
Txio pertsonalak	4.046.512	1.128.124
Birtxioak	4.345.500	963.668
Euskarazko txio pertsonalak	2.634.534	530.226
Euskarazko birtxioak	2.421.058	400.448

3.3. Txiolarien harreman sareak

Atal honetan euskal txiolarien harremanak zeintzuk diren aztertzeari ekin zaio, birtxioetan oinarrituta. Garatutako gazte/heldu identifikadoreari esker, iragarritako gazte eta helduen harreman sareak argitaratzeko aukera edukiko da, helduekin eta gazteekin talde bana osatuz eta talde bakoitzaren harreman sarea modu independentean landuz. Harreman-sareak nola josten diren ikustarazteko, euskarazko ia 3 milioi birtxio erabiliko dira, 400.448 gazteei dagozkienak eta 2.421.058 helduei dagozkienak. Elkarrekintzen sare bat lortuko da horrela, birtxio bakoitza elkarrekintza bat izanik.

Grafoa sortzeko, birtxio bakoitzetik bi datu erabili dira, alde batetik, zeinek birtxiokatu duen eta, bestetik, zein izan den birtxiokatua. Bi datu horietatik abiatuta sare bana eraiki da, heldu zein gazteentzat, *gephi* programa (Bastian *et al.*, 2009) baliatuz. Lehenik eta behin, sare handi hauek azpitaldeetan zatitu dira modularitatea erabiliz (Blondel *et al.*, 2008), sare bakoitzean komunitateak antzemateko asmoarekin. Bigarren pausoa, sareari egitura espaziala ematea izango da, erlazionatutako nodoei zentzuzko egitura bat emanez, horretarako *ForceAtlas2* algoritmoa erabiliz (Jacomy *et al.*, 2014). Azkenik, azpi-talde edo komunitate bakoitzaren identitatea definitu da, horretarako azpi-talde hauetako nodo garrantzitsuenak erabili direlarik. Era honetan, grafo bakoitzaren komunitateak nola eta zeren inguruan egituratzen diren ikusi da, alde batetik 33.277 nodo edo erabiltzailez osatutako helduen grafoa eta, bestetik, 24.987 nodo edo erabiltzailez osatutako gazteen grafoa aztertuz.

Helduen grafotik eratorritako azpitaldeei erreparatzen badiegu (3. taula eta 1 irudia), ikusi daiteke gai nahiko ezberdinen inguruan harremanak direla. Ezberdintasunak ezberdintasun, ikusi daiteke talde ezberdinek komunean daukaten ezaugarria, talde guztiek Euskal Herriko gaiekin erlazioa daukatela. Ikusi daiteke, helduen kasuan, euskal munduaren inguru hurbilari buruz hitz egiteko erabiltzen dela euskara Twitterren. Hau da, helduek egunerokotasuna komentatzeko kanal moduan hartzen dute euskarazko Twitter, azpitalde ezberdinak gertaera hurbilekin erlazionatuta daudelarik. Honela, nahiko modu errazean erlazionatu ahal izan dira taldeak gai zehatzekin, jarraian ikusi ahal izango den moduan.

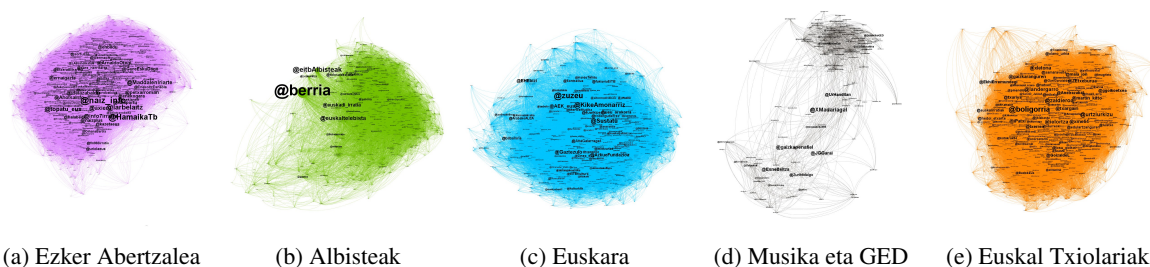
¹Argibide gehiago nahi izatekotan ikusi Master Amaierako Lana (Fernández de Landa, 2018) <https://addi.ehu.es/handle/10810/29605>

- *Ezker Abertzalea* (nodoen % 27,92): Azpitalde hau orientazio politiko konkretu bat daukaten pertsonen nodoez osatuta dago, zehazki Ezker Abertzalea osatzen duten erabiltzaileengatik osatuta dago. 3. taulako lehenengo zutabearen agertzen diren erabiltzaileak alde batera utzita, badaude nodo garrantzitsu ugari ezker abertzaleko militante edo erakundeak direnak (@ArnaldoOtegi, @sortuEH, @jpermach, @LABsindikatu, @JosebaAlvarez...), nodo garrantzitsuenek baino hobeto azaltzen dutelarik talde honen idiosinkrasia.
- *Albisteak* (nodoen % 23,77): Ia erabiltzaile guztien laurdena batzen duen bigarren talde hau albisteekin erlazionatuta egongo litzateke. Erabiltzaile gehienak komunikabideekin erlazionatuta daude, besteak beste, EITB taldeko hainbat erabiltzaile aurkitu daitezkeelarik.
- *Euskara* (nodoen % 15,34): 3. azpitaldean, euskararekin erlazionatutako gaiak daude, euskarazko komunikabideak (@zuzeu, @Gaztezulo, @ArabakoALEA), euskara sustatzeko elkarteak (@AEK_eus, @EHEbizi...) eta baita euskararekin erlazionatutako hainbat norbanako (@KikeAmonarriz, @KoldoTellitu, @MertxeMugika) azaltzen direlarik.
- *Musika eta GED* (nodoen % 13,56): 4. azpitalde honetan, fenomeno berezi bat ematen da, dena talde berdinean bateratua egon arren, bi taldetxo ezberdinek osatzen baitute azpitalde hau. Era honetan, nahiz eta dena bateratua azaldu, bi buru dituela esan daiteke. Lehenengo zatia musikari buruzkoa izango litzateke, musika talde asko bertan baitaude (@EsneBeltza, @ZuriHidalgo, @ZeEsatek, @40minuturock, @hesiantaldea, @ItzrrSemeak...), beraz, musikarekin erlazionatutako taldetxoa izango genuke. Bigarren zatia osera Gure Esku Dago dinamikarekin zerikusia daukaten erabiltzaileekin erlazionatuta egongo litzateke (@GoierrikoGED, @GEDTolosaldea, @GureEskuDagoDon...).
- *Euskal Txiolariak* (nodoen % 13,10): Azken azpitalde honetan, erabiltzaile euskaldun ohikoak edukiko genituzke, euskal komunitatearen artean garrantzitsuak diren txiolariak izanik. Norbanako ezagunez osatutako taldea izango litzateke, euskal komunitatean jarraituak izanik.

3. taula. Helduen azpitalde bakoitzeko nodo garrantzitsuenak.

Ezker Abertz.	Albisteak	Euskara	Musika eta GED	Euskal Txio.
@naiz.info	@berria	@zuzeu	@XMadariagaI	@boligorria
@HamaikaTb	@eitbAlbisteak	@KikeAmonarriz	@gaizkapenafiel	@zaldieroa
@larbelaitz	@euskaltelebista	@Sustatu	@JGGarai	@urtziurkizu
@topatu_eus	@euskadi_irratia	@Gaztezulo	@EsneBeltza	@landergarro
@axierL	@tolosaldeataria	@AEK_eus	@UrHanditan	@ielortza

1. irudia. Helduen Harremanak Twitterren



Gazteen grafotik eratorritako azpitaldeei erreparatu gero (4. taula eta 2 irudia), ikusi daiteke, antzekotasun eta ezberdintasunak daudela helduen grafoarekin konparatzerakoan. Helduen grafoarekiko antzekotasunei erreparatu, ikusi daiteke, *Euskara*, *Ezker Abertzalea* zein *Albisteak* taldeak bi grafoetan azaldu direla. Bi grafoetan azalduetako azpitalde hauek politika (*Ezker Abertzalea*) eta berehalakotasunarekin (*Albisteak*) erlazionatu ditzakegu, Twitterren identitatearen oinarritzko ezaugarriak direnak. Bestetik, ezberdintasunei so eginez gero, erreparatu daiteke

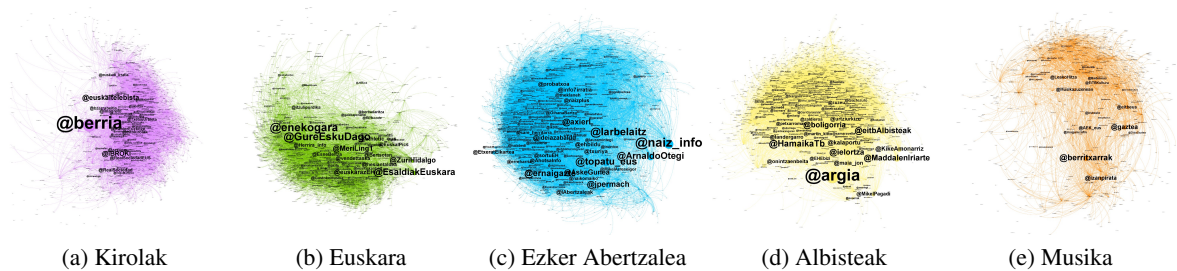
aisialdiarekin lotutako gaiek badaukatela bere pisua gazteen harremantzeko moduetan, *Kirola* eta *Musika* gai garrantzitsu moduan azaltzen baitira, erabiltzaileen herena osatzen dutelarik bi taldeen artean. Helduekiko antzekotasun eta ezberdintasunak argitu ostean, azpimarratu beharra dago, kasu honetan ere egunerokotasuna komentatzeko kanal moduan hartzen dutela euskarazko Twitter, azpialde ezberdinak gertaera hurbilekin erlazionatuta daudelarik kasu honetan ere.

- *Kirolak* (nodoen % 21,61): Gazteen ereduko nodo gehienak barnebiltzen dituen azpialde hau, kirolen ingurukoa da, bertako kirol talde edo erakundeez (@RealSociedad, @RealSociedadEUS, @ASPEpelota, @SDEibar...) zein bertako kirolariez (@InigoMartinez, @AmetsTxurruka, @XabierUsabiaga, @Markelirizar...) osatuta baitago. Hala ere, horiek guztiak baino garrantzitsuagoak dira, nodo tamainagatik, kirol kazetariak (@iBROKI, @XabierEuzkitze, @Imagreto, @bzarrabeitia, @TxetxuUrbietia...) zein komunikabideak (@berria, @euskaltelebista, @eitbkirolak, @euskadi_irratia...). Beste behin ere, argi ikusi daiteke, kirolen taldea izan arren, komunikabideek badutela protagonismo erraldoia.
- *Euskara* (nodoen % 20,70): Bigarren azpialde honetan, nodo guztien bosten bat edukiko genuke, talde handia izanik. Kasu honetan, nodo garrantzitsuenak euskararen arloarekin zerikusia daukatela ikusi daiteke (@EsaldiakEuskara, @euskarazEH, @Bertsotan, @bertsolaritza, @Euskeraz_Bizi...). Helduen azpialdeetan ere gai honekin erlazionatutako komunitate bat aurkitu da, nahiz eta nodoak ezberdinak izan, gaia euskararekin erlazionatuta dagoela esan daiteke.
- *Ezker Abertzalea* (nodoen % 17,12): Hirugarren talde hau Ezker Abertzalearekin erlazionatutako nodoez osatuta egongo litzateke. Besteak beste, Ezker Abertzalearen orbitako komunikabide (@naiz_info, @topatu_eus, @info7irratia, @naizplus...) erakunde (@ernaigazte, @ehbildu, @sortuEH...) zein norbanakoez (@ArnaldoOtegi, @lauramintegi...) azpialdeari identitatea ematen diotelarik. Beste behin ere, ikusi daiteke helduen azpialdeekiko antzekotasunak daudela, kasu zehatz honetan ere oso antzekoak direlarik.
- *Albisteak* (nodoen % 14,92): 4. talde honetan albisteekin erlazionatu ditugu bertako nodoak, nodo deigarrienak euskal komunikabideekin zerikusia baitaukate (@argia, @HamaikaTb, @eitbAlbisteak, @zuzeu, @Gaztezulo). Kasu honetan ere, azpialde hau eta helduen ereduaz azaldutako azpialdeak erlazioa daukate, bi eredu ezberdinen artean paralelotasunak ematen direla berriz ere erakutsiz.
- *Musika* (nodoen % 11,35): Azkeneko azpialdea, nahiko talde anitza izan arren, musikarekin erlazionatutako aktore ezberdinekin erlazionatu ditzakegu nodo garrantzitsuenetako batzuk. Aktore horien artean musika hizpide duten komunikabideak (@gaztea, @DidaGaztea), musika taldeak (@berritxarrak, @muguruzafm, @Glaukomaband) eta baita disketxeren bat ere bai (@BagaBigaeus) edukiko genituzke.

4. taula. Gazteen azpialde bakoitzeko nodo garrantzitsuenak.

Kirolak	Euskara	Ezker Abertz.	Albisteak	Musika
@berria	@enekogara	@naiz_info	@argia	@berritxarrak
@euskaltelebista	@GureEskuDago	@larbelaitz	@HamaikaTb	@gaztea
@iBROKI	@EsaldiakEuskara	@topatu_eus	@eitbAlbisteak	@izanpirata
@RealSociedad	@ZuriHidalgo	@ArnaldoOtegi	@MaddalenIriarte	@eitbeus
@XabierEuzkitze	@MeriLing1	@ernaigazte	@ielortza	@LeakoHitza

2. irudia. Gazteen Harremanak Twitterren



Behin bi ereduetako azpi taldeak interpretatu eta definitu direnean, bien arteko konparaketa egin da. Ikusi daiteke, bai helduek eta baita gazteek ere, harremantzeko orduan preferentzia ezberdinak dituzten arren, parte komun bat daukatela. Alde batetik **sare sozial honen izaera politikoa eta berehalakotasuna** izango litzateke parte komunaren zati bat, nahiz eta helduen kasuan nabariago izan. Gainera, aipatu beharra dago, azpitalde ia gehienetan komunikabideekin erlazionatutako erabiltzaileek garrantzizko papera jokatzeko dutela, Twitterren izaera informatiboa konfirmatzen delarik. Bestalde, azpimarratu beharra dago, **erabiltzaile euskaldunek euskara erabiltzen dutela, batez ere, beren inguruko gertakari edo gaien inguruan hitz egiteko**. Amaitzeko, esan beharra dago, euskal txiolarien komunitateak Twitterren ezaugarri orokorra konpartitzeaz gain, hau da, berehalakotasuna eta izaera politikoa, baduela berezitasun propio bat, euskara erabiltzearena euskaldunen kontuez aritzeko.

4. Ondorioak

Twitter sare sozialera konektatuta dauden euskal hiztun gazte eta helduen on-line harremanetara hurbilpen bat egitea lortu da. Helburu orokor hori betetzeko hainbat pausu eman dira, lehenik eta behin Twitter sare sozialetik euskal erabiltzaileen datu kantitate erraldoiak erauzi dira. Bigarrenik, erauzitako euskal txiolaria sailkatu dira gazte eta heldu artean. Hirugarrenik eta azkenik, gazte eta helduen harremanak nola ematen diren argitu da, bi talde ezberdinen errealitatea zein den erakutsiz. Lan honekin, frogatuta geratzen da gizarte-zientzia eta konputazio-zientzien arteko konbinaketa aberasgarria dela, Hizkuntzaren Prozesamenduko teknikak aplikatuz, ezaugarri demografikoak iradoki edota komunitateen azterketa bezalako atazak burutu daitezkeela erakutsi delarik.

Datuen erauzketaren inguruko balorazioa burutzerakoan frogatu da Twitter datu-iturri garrantzitsu bat dela, informazioa masiboki eta kostu txikiarekin lortzeko aukera ematen duena. Erauzketa teknika honi esker, datu kantitate handiak (Big Data) lortu dira oso kostu txikiarekin ia 8.000 euskal erabiltzaile erauztea lortu baitira, 10 milioi txio baino gehiago geureganatuz. Gainera, datu mordoaren artean ere euskaldunena bezalako komunitateak aurkitu daitezkeela ikusi da.

Ikerketa lan honen atal garrantzitsuetako bat gazteak eta helduak ezberdintzean oinarritu da, interesetako bat gazteen errealitatea ezagutzea izanik. Gazte/heldu ezberdintzea adinaren arabera burutu ordez, testuaren formaltasunaren arabera burutu da, adinaren arabera etiketatzeak kostu altuegia baitzuen. Hortaz, adinaren etiketatzearen zailtasunen aurrean, lasterbide metodologikoa erabiltzea erabaki da, testu informalaren kontzentrazioa altua bada gaztea izango dela erabakiz. Honela, txioen corpus txiki bat etiketatu ostean, sailkapena burutu da *IXA pipes* dokumentu sailkatzailea erabiliz erabiltzaile bakoitzaren testuaren formaltasunean oinarrituta. Sistema honek, emaitza fidagarriak lortzeaz gain, etiketatutako corpus txikiarekin ere sailkapena ondo egiten duela esan beharra dago. Testuaren oinarrituta ezaugarri demografikoak iragartzea posiblea dela frogatuz, baita euskal hiztunen komunitatean ere.

Behin gazte eta helduen taldeak ezberdinduta, harremanak nola ematen diren azaleratu da, horretarako euskarazko birtxioetan oinarrituta harreman sare bat sortu delarik. Honez gain, aipatu beharra dago, komunikabideekin eta politikarekin (Ezker Abertzalea) erlazionatutako erabiltzaileek garrantzizko papera jokatzeko dutela, Twitterren izaera informatiboa konfirmatzen delarik. Horrez gain, ondorioztatuta daiteke, euskara, euskaldunen gaiei buruz hitz egiteko erabiltzen dela gehienbat.

Lan honek euskara XXI. mendeko testuingurura nola moldatzen ari den ezagutzeko aukera eman du ere, ikusi ahal izan da, euskara presente dagoela sare sozialetan, euskarazko 6 milioi txio baino gehiago lortu baititugu ia 8.000 erabiltzaile ezberdinenak. Euskara sare sozialak bezalako teknologia berrietan presente egoteak esan nahi du, geure hizkuntza erronka berrietara egokitzeko kapaza dela, euskaldunen sorkuntza kapazitateari esker

teknologia berrietan ere erabilia delarik. Sarean euskara euskaldunen gaiei buruz aritzeko erabiltzeak erakusten digu, globalizatutako eta etengabe konektatutako mundu honetan ere, euskaldunek badutela gaitasun berezi bat beren lekua bilatu eta bertan finkatzeko.

5. Etorkizunerako planteatzen den norabidea

Etorkizunera begira, lan honetan garatutako metodologia bera aplikatu daiteke beste motatako komunitateetan, erabakitzen den komunitatearen baitan adinaren arabera banaketa egin ahalko litzateke eta barneko harremanak aztertu. Honek erakusten du, egindako lana ez dela mugatua eta aplikazio ugari izan ditzaken metodologia batean oinarritzen dela. Horrez gain ere, idazketa formatik habiatuta beste aldagai demografiko batzuk iragartzea izango da asmoa, generoa edota klase soziala antzemateko aukera eduki arte, dena soziolinguistika eredu hartuta.

6. Erreferentziak

- Agerri, Rodrigo, eta German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence* 238.63–82.
- Al Zamal, Faiyaz, Wendy Liu, eta Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. *ICWSM* 270.2012.
- Bastian, Mathieu, Sebastien Heymann, Mathieu Jacomy, eta others. 2009. Gephi: an open source software for exploring and manipulating networks. *Icwsim* 8.361–362.
- Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, eta Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008.P10008.
- Conover, Michael D, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, eta Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.
- Fernández de Landa, Joseba. 2018. Gazteak eta euskara sare sozialetan. zer, nori, nork: euskarazko txio formal eta informalak sailkatuz eta konparatuz.
- Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, eta Mathieu Bastian. 2014. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one* 9.e98679.
- Morgan-Lopez, Antonio A, Annice E Kim, Robert F Chew, eta Paul Ruddle. 2017. Predicting age groups of twitter users based on language and metadata features. *PloS one* 12.e0183537.
- Nguyen, Dong, A Seza Doğruöz, Carolyn P Rosé, eta Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics* 42.537–593.
- , Rilana Gravel, Dolf Trieschnigg, eta Theo Meder. 2013. "how old do you think i am?" a study of language and age in twitter. In *ICWSM*.
- Rao, Delip, David Yarowsky, Abhishek Shreevats, eta Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37–44. ACM.
- Roesslein, Joshua. 2009. tweepy documentation. *Online*] <http://tweepy.readthedocs.io/en/v3.5>.
- Rosenthal, Sara, eta Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 763–772. Association for Computational Linguistics.
- Umap.eus, 2018. Twitterreko euskarazko jarduna.
- Zubiaga, Arkaitz, Bo Wang, Maria Liakata, eta Rob Procter. 2017. Stance classification of social media users in independence movements. *Catalonia* 2.10–960.

7. Eskerrak eta oharrak

Lan hau Hizkuntzaren Analisia eta Prozesamendua masterreko amaierako proiektuaren zati bat da (Fernández de Landa, 2018), IXA taldearen laguntzaz eta Eusko Ikaskuntzak diruz lagundutakoa.