Gazteak eta euskara sare sozialetan. Zer, nori, nork:
euskarazko txio formal eta informalak sailkatuz eta konparatuz[1]

Fernandez de Landa, Joseba

joseba.fdl@gmail.com

Twitter sare sozialetik erauzitako euskarazko 6 milioi txioak abiapuntutzat hartuta, euskal hiztun gazte eta helduen inguruko ikerketa burutu da, hauek nola erlazionatzen diren eta zertaz aritzen diren erakutsiz. Prozesu hori burutzeko, metodologia propioa garatu da, batetik Lengoaia Naturalaren Prozesamenduko teknologiak baliatuz gazte eta helduak sailkatzen dituen sistema inplementatu da, eta bestetik, erabiltzaileen gaiak iradoki eta harremanak identifikatzen dituzten teknikak garatu dira. Guzti honekin, ikerketa sozialak eta konputazio zientziak bateragarriak direla erakutsi da, ikerketa egiteko modu berriei bide emanez.

Giltza-hitzak: Sare sozialak. Euskara. Gazteak. Gaiak. Harremanak. Soziologia konputazionala.

1. Sarrera[2]

Sare sozialak adierazpen librerako esparru bilakatu dira, nahi dena nahi den momentuan esateko aukera zabalduz eta gazteen esparru bezala kontsideratuz. Honek aukera ematen du euskararen inguruko hausnarketa burutzeko, batez ere gazteen errealitatea ezagutzeko. Euskara teknologia berrietara nola moldatzen den jakiteko eta gazteen euskararekiko atxikimendua ezagutzeko asmo bikoitzarekin, sare sozialetan euskararen inguruko ikerketa burutu da, zehazki Twitterren. Teknika berritzaileetan oinarritutako ikerketa honi esker, sare sozial honetan garrantzitsuenak diren euskal erabiltzaileen komunitate eta gaiak identifikatzea lortu da. Gainera, gazteak helduengandik bereizi dira, gazteengan bereziki fokua jartzeko intentzioarekin, haiek izango baitira euskararen etorkizunaren bermea. Alde batetik, euskara ze gaietan indartsu dagoen azaleratuko da, euskal txiolarien ohiko gaiak edo lehentasunak zeintzuk diren ezagutzeko. Bestalde, txiolari euskaldunen harremanetan oinarrituta, hauen azpitaldeak ezagutuko dira, euskal txiolariak nola harremantzen diren modu sakonean ezagutzeko. Horretarako, datu kopuru handiak (*Big Data*) erauzi dira Twitterretik eta hauek Ikasketa Automatikoan (*Machine Learning*) oinarritutako tekniken bitartez arakatuak izan dira. Era honetan, euskal gazteen komunitatearen baitan, euskararen

indar-guneak azaleratuko dira, gazteak motibatzen dituzten gaiak ezagutaraziz eta zein harreman sareetan mugitzen diren erakutsiz. Guzti honen intentzioa, euskararen aldeko edozein lanketan gazteen ikuspegia txertatzen laguntzea izango da, gazteen lehentasunak azaleraziz.

2. Proiektuaren definizioa

Lan honen asmoa, gazteak eta euskara aztertzea izango da sare sozialetan. Era honetan, gazteen errealitate ezezagunera hurbilpen bat lortzeaz gain, XXI. mendeko erronketara euskara nola egokitzen ari den ezagutu ahalko da. Honela, Twitter sare sozialean gazteek zertaz eta zeinekin aritzen diren azaltzea izango da intentzioa. Euskal txiolari gazteen errealitatea aztertuko duen ikerketa lan hau burutzeko, lana lau 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, bereziki gazteen errealitatea ezagutzea interesatzen baitzaigu. Hirugarrenik, euskal txiolarien gaiak identifikatzeko, txioen testuan oinarrituko gara, bertatik gaiak ondorioztatzeko. Laugarrenez eta azkenik, euskal txiolarien harremanak ezagutzeko, euskarazko birtxioetan oinarrituko gara, hauekin harremanen sare bat sortuz. Izendatu den ataza zehatzaren arabera metodologia zehatz bat erabiliko da, jarraian ikusi daitekeen moduan.

Datuen erauzketa: Datuak erauzi aurretik, ikertuko den unibertsoa zehaztu da, kasu honetan euskal txiolariak direlarik. Euskal erabiltzaileak Twitterreko sarean identifikatzeko, Umap.eus web orriaren zerrenda erabiliko da, webgune honek euskal txiolariak identifikatzen baititu Twitterren (Umap.eus, 2018). Behin ikertu beharreko unibertsoa zein den argi edukita, bertatik datuak erauzteari ekin behar zaio. Erauzketarako Pythoneko *tweepy* paketea erabili da.

Gazte/heldu sailkatzailea: Atal honen helburua, euskal txiolari gazteen identifikazioa izango da, txiolariak gazte eta heldu artean sailkatuko direlarik. Sailkapen hau, ez da adinaren arabera egingo, bizitza etaparen arabera baizik, eta horretarako txio pertsonalen testua hartuko da erreferentziatzat. Txioen testuan oinarrituta, txio bakoitza formal edo informal moduan sailkatua izango da, txio informalen kontzentrazioa handia denean gaztea dela ondorioztatuz, gazteagoek estilo ezohikoagoa edo informalagoa daukatela (Nguyen et al., 2014) oinarri izanik. Txioak formal eta informal artean bereiziko dituen sailkatzailea entrenatzeko *IXA pipes* dokumentu sailkatzailea erabili da (Agerri eta Rigau, 2016).

Txiolarien ohiko gaiak identifikatu: Datu moduan euskarazko txio pertsonalak erabili dira gaien identifikaziorako. Horretarako, Topic-Modeling teknika erabili da, LDA algoritmoa aplikatuz, *gensim* paketea (Rehurek eta Sojka, 2010) erabili. Bistaratze intuitibo bat lortzeko, emaitzak irudi moduan argitaratzen dituen *LDAvis* metodoa (Sievert eta Shirley, 2014) erabili da.

Txiolarien harremanak azaleratu: Datu moduan euskarazko birtxioak erabili dira harremanak zeintzuk diren iradokitzeko. Horretarako, nodo eta beren loturen arteko grafo bat eraikiz eta azpi-komunitateak antzemanez. Grafoaren irudikapenerako *Ghepi* programa (Bastian et al., 2009) erabili da, nodoak eta beren loturak azalduko dizkiguna. Horrez gain, grafo barneko azpi-komunitateak azaleratzeko modularitatea (Blondel et al., 2008) erabiliko da.

3. Datuen erauzketa

Lehenik eta behin, euskal txiolariak identifikatzean oinarritu gara da, Twitterren euskara erabiltzen duten erabiltzaileak hain zuzen. Horrela, euskal erabiltzaileak hautemateko, euskarazko txio kopuru minimo bat duten erabiltzaileak hautatuko dira. Euskarazko twitter erabiltzaileak lortzeko, umap.eus webguneko euskal txiolarien zerrenda erabili da (Umap.eus, 2018). 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). Honekin batera, zerrendako erabiltzaile hauek aktibo egon behar dira azkeneko hilabeteetan. Era honetan, euskal txiolarien zerrenda lortu da, 8.189 erabiltzaile euskaldunek osatzen dutena.

Behin erauzi nahi den unibertsoa definituta dagoelarik, bertatik informazioa erauzten hasteko prest gaude. Datuen erauzketa burutu ahal izateko Twitterreko APIa erabili da, honetarako Pythoneko *tweepy* paketea hautatuz. Datuen bilketa hau burutu ahal izateko erabiltzaile bakoitzak publikatutako

txioak erauzten dira. Twitterreko APIak 3.200 txioko muga du, beraz erabiltzaile bakoitzeko gehienez txio kopuru hori lortu ahal izango da. Muga honetaz gain ere, denbora muga bat gehitu behar zaio APIari, 15 minuturo soilik 15 erabiltzaile erauzi ahal dira, erauzketa asko luzatuz denboran zehar.

Datuen lorpenerako teknika honi esker, 2018ko maiatzaren 30 eta 31an burututako erauzketan, 7980 euskal erabiltzaile erauztea lortu da, 10 milioi txio baino gehiago geureganatuz oso kostu material txikiarekin. Honela 5.198.043 txio pertsonal lortu dira, horietako 3.171.485 (% 61) euskaraz eman direlarik. Bestalde, 5.473.031 birtxio lortu dira ere, horietako 2.891.136 (% 53) euskaraz izanik.

4. Gazteak identifikatzea

Erauzitako erabiltzaileen artean gazteak eta helduak ezberdintzeko asmoarekin hainbat ikerketa ezberdin aztertu dira, Twitterreko erabiltzaileen adina iragartzen dituzten sistemak kontutan hartuz. Aztertutako sistema guztiek komunean daukate adin tarte konkretuen arabera sailkatzen dutela erabiltzailea. Hala ere, artearen egoerako sistema onena, adin tarteekin baino, bizitza etapekin sailkapen hobea egiten duela ikusi da (Nguyen et al., 2014), denboran zeharreko esperientzia konpartituak argiago ikusten baitira bizitza etapetan (Nguyen et al., 2016; Eckert, 2017). Horregatik, adin tarte zehatzetan zentratu ordez, bizitza etapetan zentratzea erabaki da sailkapena egiteko, gazte/heldu klaseak aukeratuz adin zehatza markatu ordez.

Bibliografiako sistema ezberdinetan ikusi ahal izan den moduan, sistema guztiak erabiltzaileen etiketatzean oinarritu dira, erabiltzailearen adina edo adin tartea eskuz etiketatu dutelarik. Erabiltzaile euskaldunen kasuan, ostera, zailtasunak aurkitu dira erabiltzaileen adina eskuz etiketatzeko, askok identitatea ezkutuan mantentzen baitute. Zailtasun honen aurrean, saihesbide metodologiko bat burutzea proposatu da, bibliografiako sistemen etiketatze estrategiatik aldendu arren, sistema hauek argitaratutako ondorioetan zentratzen dena. Hau da, adina iradokitzeko ezaugarririk garrantzitsuena idazkeran oinarritzen da (Rao et al., 2010; Al Zamal et al., 2012; Nguyen et al., 2014; Morgan-Lopez et al., 2017) eta horretan zentratzea erabaki da. Bibliografiako sistemen ondorioetan antzeman den moduan, helduek gazteak baino hitz konbentzionalagoak erabiltzen dituzte (Nguyen et al., 2014), 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 informalagoa daukatelarik (Nguyen et al., 2014). 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 etiketatu ordez, txioak etiketatzeari ekingo zaio, formal eta informal artean desberdinduz, zailtasunak gaindituz. Era honetan, txioen motaren kontzentrazioren arabera sailkatuko dira erabiltzaileak gazte eta heldu artean.

Gazte eta heldu klaseak iragartzeko, txio pertsonal bakoitzaren testuan oinarrituko da gure sailkatzailea, testu hau segun eta nola idatzia izan den, heldu edo gazte moduan etiketatuko duelarik. Era honetan, testuaren estiloaren arabera bi testu mota ezberdinduko ditugu, formalak eta informalak. Erabiltzaileak testu motaren arabera sailkatuko direnez, txio formal edo informalen kopuruaren arabera determinatuko da gazteak edo helduak diren. Honela, erabiltzaile baten txioen gehiengoa informala denean, erabiltzailea gaztea dela kontsideratuko da, txioen gehiengoa formala denean ostera, erabiltzailea heldu bezala sailkatuko da. Era honetan, erabiltzaileak ez dira itsura fisiko edo adinaren arabera sailkatuko, idazteko eragatik antzeman daitekeen izaeraren arabera baizik. Kasu zehatz honetan, euskal erabiltzaileetan zentratuko garenez, erabaki da euskarazko testuetan soilik zentratzea, hizkuntza bakoitzak sistema propio bat beharko lukeelako.

Hurrengo pausua, erabiltzaileak sailkatuko dituen sistema garatzea izango da, bizitza etaparen arabera gazte edo heldu sailkatuz eta horretarako testuan oinarrituz. Helburu hori betetzeko, *IXA pipes* dokumentu sailkatzailea erabiliko da, etiketatutako corpus txikietarako aukera ona delarik. Sistemaren entrenamendurako 1.000 txioko corpus txiki bat eskuz anotatua izan da, idazteko moduaren arabera sailkatuak izanik. Era honetan, txio batek *formal* etiketa eramango du, lengoaia estandarrean idatzia izan bada, edo *informal* etiketa, txioa modu kolokialean idatzia izan denean. Era honetan,

erabiltzailea gaztetzat hartuko da, txioen kopuru handi bat modu informalean idatzi baldin badu.

Behin sistema aplikatuta erauzitako 7.980 euskal txiolariengandik, 7.087 erabiltzaile sailkatzea lortu da, gutxienez 10 txio euskaraz dauzkaten erabiltzaileak soilik sailkatu direlarik, besteak kanpo utziz. Sailkatzaileak txioak informal eta formal artean sailkatzen dituenez, txio informalen kontzentrazioan oinarrituko dugu erabiltzaile baten gaztetasuna. Era honetan, erabiltzaile gazte moduan 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 informalen bitartez erabiltzaile gazteak antzemanez, adina igartzeko zailtasuna alboratuz eta lasterbide metodologikoa aplikatuz.

Era honetan 5.508 erabiltzaile heldu bezala kontsideratu dira eta 1.579 gazte moduan, beti ere idazteko era kontutan hartuz. Gazte eta helduen erabiltzaileen banaketa nahiko desorekatua dagoela esan daiteke, gazteak nabarmen urriago izanik. Sailkatutako erabiltzaileen txio kopuruari erreparatuz gero, 10 milioi txioetatik, 8 milioi helduen taldeari dagozkio eta 2 milioi gazteen taldeari, berriz ere desoreka dagoela antzemanez. Hala ere, talde bakoitza modu independientean aztertuko denez, tamainaren arabera moldatu beharko dira aplikatuko diren ikerketa teknika ezberdinak.

5. Txiolarien ohiko gaiak identifikatu

Euskal txiolarien gaiak azaleratzeko, euskarazko 3 miloi txio pertsonal baino gehiago erabiliko dira, txio hauen testuetan oinarrituz gaiak azaleratuz. Helburu horretarako, testu meatzaritzan maiz erabiltzen den *Topic modeling* tresna erabiliko da. Tresna honekin, hitzak sailkatzeari ekingo zaio, antzeko hitzekin topiko orokorrago bat sortuz. Beste hitz batzuekin esanda, hitzak clusterizatuko dira, topiko bakoitzarekin zerikusia daukaten hitzak taldekatuko dituena. Lan honetan, hitzen taldekatzeari esker, euskal txiolariek zein gaiei buruz hitz egiten duten identifikatzen saiatuko gara. Hitzen sailkapen hau egiteko *Latent Dirichlet Allocation* (LDA) teknika erabiliko da, testuinguru bereko hitzekin topikoak sortuz.

LDA, testu corpusak bezalako datu diskretuen bildumetan aplikatzeko eredu probabilistiko generatibo bat dugu (Blei et al., 2003). Eredu estatistiko generatibo honek, ahalbidetu egiten du behagarriak diren gertakizunak, latente edo ezkutuan dauden multzoetan sailkatzea. Sailkapen hau, ezkutuan dauden hainbat ezaugarriei esker gertatzen da, adibidez, hitzen antzekotasuna. Honela, hitzak eta dokumentuak erlazionatuz, ezkutuan dauden erlazioak azaleratuko dira, topiko bezala ezagutuko ditugun kategorietan. LDA algoritmoak ondo funtzionatzeko, erabiltzaile bakoitzaren txioekin sortutako dokumentuetan aplikatuko dugu (Hong et al., 2010; Zhao et al., 2011). Hortaz, nahiz eta txioak laburrak izan, LDA era egokian aplikatzea lortu da, erabiltzaile bakoitzaren euskarazko txio pertsonalekin dokumentu bana sortuz.

Gazte eta helduen ereduetarako topiko kopuru egokia aukeratzeko asmoarekin, topiko kopuru ezberdinekin frogak egin dira. Aipatzekoa da ez dagoela topiko kopuru 'zuzen' bat (Binkley et al., 2014), baina argi eduki behar da topiko kopuruak hauen interpretagarritasuna baldintzatuko duela (Steyvers eta Griffiths, 2007). Erabili den topikoen zenbatekoa interpretagarritasunean zein clusterren sakabanaketan oinarritu da. Errealitate sozialarekiko koherentzia edukitzea eta eredu sakabanatuena lortzea izan da asmoa. Honela, eredu bakoitzerako topiko kopuru ezberdinen arteko konparaketa burutu ostean, *helduen* eredurako 20 topiko erabiltzea erabaki da eta *gazteen* eredurako 12 topiko. Bistaraketa intuitiboago bat lortzeko asmoarekin, irudi grafikoak ematen dituen *LDAvis* metodoa erabili da, interpretazioan eta bistaratze lanetan oso lagungarria dena (Sievert eta Shirley, 2014). Topiko bakoitzaren identitatea, bere barneko hitz probableenek zehaztuko dute (Binkley et al., 2014), honi esker, euskal txiolarien gaiak zeintzuk diren argituz.

Helduen gaiei erreparatzen bazaie, gai asko politikarekin erlazionatuta daudela baiezta daiteke, adibidez, *politika*, *euskal presoak*, *herri mugimendua* (*HM*), *hezkuntza* eta *euskara*. Honek erakusten du, Twitterreko euskal erabiltzaile helduek politikarekiko grina dutela. Esan beharra dago instituzio politiko publikoek ere badutela bere tokia sare sozial honetan, besteak beste *eskaintza kultural instituzionala*, *administrazio publikoa* edota herrialdeetako udalei (*gipuzkoa*, *bizkaia* eta *nafarroa*) buruz dagokienean. Beraz, gai politikoez edota gizarte gaiez aritzeko erabiltzen dute helduek sare sozial hau, norbanako zein erakunde publikoei buruz arituz. Gizarte gaiei buruz aritzea helduen ezaugarri gisa hartu dezakegu emaitzak aztertu ostean, interes soziala duten gaiak baitira talde honetan aipatu diren gehienak.

Gazteen gaiei erreparatuz, ikusi daiteke gairik ohikoenak erlazioa daukatela norberaren bizitza propioarekin, *bizitza kontatu*, *sentimenduak azalerazi* eta *zoriondu* baitira erabiltzaile gazteen gai garrantzitsuenetarikoak, guztiak beren egunerokotasuneko gertaerekin erlazionatuta. Bestalde, euskalkietan modu kolokialean hitz egiterako orduan ere (*gipuzkera* eta *bizkaiera*), egunerokotasuneko gauzez aritu ohi dira, baina beren euskalki propioa erabiliz. Bestalde, ezin da ahaztu ere kirolak baduela bere tokia erabiltzaile gazteen artean (*kirola*, *Athletic* eta *arrauna*). Hortaz, gazteen gairik ohikoenak ingurukoekin komunikatzean oinarritzen dira, kirola presente egonez baina bigarren plano batean geratuz. Gazteetan zentratuz beraz, esan beharra dago, gazteek sare sozialetan euskara erabiltzen dutela beren eguneroko bizitzan besteekin komunikatzeko. Honekin ondoriozta daiteke, eguneroko bizitzan euskara erabiltzearen ondorio gisa, sare sozialetan ere euskara erabiltzeko joera naturala daukatela erabiltzaile hauek.

Bi adin taldeen arteko konparaketa eginez, aipatu beharra dago gazteek egunerokotasuneko gauzetarako (hitz egin, bizitza kontatu, sentimenduak adierazi, zoriondu...) erabiltzen dutela sare sozial hau. Helduek ostera, gizartean pil-pilean dauden gaiei buruz aritzen dira gehiago (politika, komunikabideak, euskara, herri mugimendua). Era honetan, esan daiteke, gazteek beren gertukoekin komunikatzeko erabiltzen dutela sare sozial hau, egunerokotasuneko gertakariak adieraziz. Helduen artean ostera, mezuak gizarteratzeko lanabes modura kontsideratzen dela ikusi daiteke, batez ere izaera politikoa daukaten gaiak plazaratzeko, komunikabideekin antzekotasun ugari erakutsiz. Twitterren berezitasunak agerian utzi dira, gazte eta helduen tematikak ezberdinak izan arren, komunikatzeko eta informazio trukerako lanabes moduan kontsideratu dezakegularik sare sozial hau. Honekin batera, aipatu beharra dago berehalakotasunean oinarritutako informazioa trukatzen dela, momentuko gertakari edo burutazioak publikatuz.

6. Txiolarien harremanak azaleratu

Laneko atal zehatz honetan, euskal txiolarien harreman sarea zein den erakustea izango da asmoa, horretarako 3 milioi birtxio baino gehiago baliatuz. Txiolarien arteko harreman sarea sortzeko birtxioak erabiltzea erabaki da, Twitterren elkarrekintza adierazteko ekintza garrantzitsuena baita. Birtxioa, beste erabiltzaile baten txio zehatz bat konpartitzean datza, erabiltzaile horrek esaten duena norberaren jarraitzaileekin konpartituz. Ondorioz, birtxiokatzeko ekintzarekin nork-nor birtxiokatu duen adierazten da, harreman zuzendu bat ezarriz, noranzko konkretu batean. Hau da, erabiltzaile batek bere atxikimendua ematen dio beste erabiltzaile batek esandakoari, harreman bat sortuz. Honela, atxikimendu asko jasotzen dituzten erabiltzaileak kontutan hartu beharko dira erreferentziazko foku bezala.

Grafoa sortzeko, birtxio bakoitzetik bi datu erabili dira, alde batetik, zeinek birtxiokatu duen eta, bestetik, zein izan den birtxiokatua. Era honetan, txio bakoitzaren abiapuntua eta helmuga izango dira gure datu iturria, erabiltzaileetan zentratuz. Horrela, pertsonetan zentratuko gara edukiak alde batera utziz, elkarrekintzaren abiapuntua eta helmuga kontutan hartuz. Bi datu horietatik habiatuta sare bana eraiki da, heldu zein gazteentzat, *gephi* programa (Bastian et al., 2009) baliatuz. Sare haundi honetan Komunitateen araberako zatiketa egiteko, modularitatea (*Modularity*) erabili da, detektatzen dituen komunitateen kalitatea ona delako eta prozesamendu abiadura azkarra delako (Blondel et al., 2008). Era horretan, birtxioen sare erraldoietik azpitalde edo komunitateak antzeman dira, erabiltzaileak birtxiokatu dituzten pertsonen arabera sailkatuz. Azpi-talde hauetan sailkatzerakoan, komunitate desberdinak zeintzuk diren ikusi ahal izango da, erabiltzaile hauek nola eta zergatik harremantzen diren azaleratuz. Taldeketari esker ere, erabiltzaileen harremanak zertan oinarritzen diren interpretatzen lagunduko digu. Honela, grafo bakoitzaren komunitateak nola eta zeren inguruan egituratzen diren ikusteko aukera emango digu atal honek, euskal txiolarien harremantzeko modu eta zergatiak argitaratuz. Honekin, euskal txiolari komunitatearen preferentziak zeintzuk diren ikusi ahal izango da, errealitate aberatsa orokortuz eta sinplifikatuz.

Helduen grafotik eratorritako azpi-taldeei erreparatzen badiegu ikusi daiteke gai nahiko ezberdinen inguruan harremantzen 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, azpi-talde ezberdinak gertaera hurbilekin erlazionatuta daudelarik. Honela, nahiko modu errazean

erlazionatu ahal izan dira taldeak gai zehatzekin, jarraian ikusi ahal izango den moduan.

Gazteen grafotik eratorritako azpi-taldeei erreparatuz gero, ikusi daiteke, antzekotasun eta ezberdintasunak daudela helduen grafoarekin konparatzerakoan. Helduen grafoarekiko antzekotasunei erreparatuz, ikusi daiteke, *Euskara*, *Ezker Abertzalea* zein *Albisteen* taldeak bi grafoetan azaldu direla. Bi grafoetan azaldutako azpi-talde hauek politika (*Ezker Abertzalea*) eta berehalakotasunarekin (*Albisteak*) erlazionatu ditzakegu, Twitterren identitatearen oinarrizko 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, azpi-talde ezberdinak gertaera hurbilekin erlazionatuta daudelarik kasu honetan ere.

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, azpi-talde ia gehienetan komunikabideekin erlazionatutako erabiltzaileek garrantzizko papera jokatzen 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. Azaldu diren harreman azpi-taldeetan ikusi ahal izan dugu, harremantzeko modua komunitate konkretuen baitan ematen dela, komunitate horien gaia nahiko erraz intuitu daitekeelarik. Harreman azpi-taldeen gaiak intuitzeari esker, jabetu gaitezke azpi-komunitate hauek batzen dituzte hari-eroaleak Euskal Herriko testuinguruarekin erlazionatuta daudela. Horregatik, ondorioztatu daiteke, euskara, euskaldunen gaiei buruz hitz egiteko erabiltzen dela gehienbat. 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.

7. Ondorioak eta etorkizuneko lana

Twitter sare sozialera konektatuta dauden euskal hiztun gazteen on-line errealitatera 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 txiolariak sailkatu dira gazte eta heldu artean, gazteen errealitatea ezagutzea baita helburu nagusia. Hirugarrenik eta azkenik, gazte eta helduen gaiak zeintzuk diren eta 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 diskurtso analisia bezalako atazak burutu daitezkeela erakutsi delarik.

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 araberako etiketatzeak kostu altuegia baitzuen. Hortaz, adinaren etiketatzearen zailtasunen aurrean, lasterbide metodologikoa erabiltzea erabaki da, testu informalaren kontzentrazioa altua bada gaztea izango dela erabakiz. Honela, txio corpus txiki bat etiketatu ostean, sailkapena burutu da *IXA pipes* dokumentu sailkatzailea erabiliz erabiltzaile bakoitzaren testuaren formaltasunean oinarrituta.

Behin gazte eta helduen taldeak ezberdinduta daudelarik, hauen gai ohikoenak zeintzuk diren azaleratu eta hauen harremantzeko modua zein den argituko da. Lehenik eta behin, euskal erabiltzaileek ze gairi buruz aritzen diren argituko da, horretarako txio pertsonalen testuetan LDA teknika aplikatzean oinarritu garelarik. Gazteek gehienbat, beren gertukoekin komunikatzeko erabiltzen dute sare sozial hau, egunerokotasuneko gertakariak adieraziz. Helduen artean ostera, mezuak gizarteratzeko lanabes modura kontsideratzen dela ikusi daiteke, batez ere izaera politikoa daukaten gaiak plazaratzeko, gizartean pil-pilean dauden gaiei buruz arituz. Gazte eta helduen tematikak ezberdinak izan arren, komunikatzeko eta informazio trukerako lanabes moduan erabilia da sare sozial hau. Bigarrenik, harremanak nola ematen

diren azaleratu da, horretarako euskarazko birtxioetan oinarrituta harreman sare bat sortu delarik. Honez gain, aipatu beharra dago, komunikabideekin erlazionatutako erabiltzaileek garrantzizko papera jokatzen dutela, Twitterren izaera informatiboa konfirmatzen delarik. Erabiltzaile euskaldunek euskara erabiltzen dute, batez ere, beren inguruko gertakari edo gaien inguruan hitz egiteko, gaien arabera ere harreman taldeak sortuz. Harremanak komunitate konkretuen baitan ematen da eta Euskal Herriko testuinguruarekin erlazionatuta daudela esan daiteke. Ondorioztatu daiteke, euskara, euskaldunen gaiei buruz hitz egiteko erabiltzen dela gehienbat.

Lan honen galdera nagusiari erantzuteko asmoarekin, euskaraz aritzen diren gazteak zertaz aritzen diren eta zeinekin harremantzen diren argituko da orain. Sarritan ezezaguna den gazteen errealitateari hurbilpen honekin, etorkizuna izango diren gazteen euskararekiko portaera ikusi ahal izango da. Gazteak euskaraz hitz egitera bultzatzen dituen gaiak eta harremantzeko moduak azaleratzea izango da asmoa, gazteak euskaraz hitz egitera animatzen dituzten zergatiak argituz. Gazteak eguneroko bizitzari eta kirolei buruz aritzen dira gehienbat, honek erakusten du, beren gertukoekin komunikatzeko erabiltzen dutela sare sozial hau, egunerokotasuneko gertakariak adieraziz. Harremanei begira, euskal erabiltzaileak intereseko gaien arabera harremantzen direla ikusi da, gazteen harremanak, euskal herriko gai politikoekin eta aisialdiarekin lotzen direlarik.

Lan honek ere, 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. Euskal komunitatearen baitako erakunde eta norbanako erreferentzial gehienak komunikabideekin eta Ezker Abertzalearekin zerikusia daukatela ikusi ahal izan da. Honez gain, erabiltzaile euskaldunek eginiko erabileratik ondorioztatu dezakegu, euskara euskaldunen inguruko gaiez aritzeko erabiltzen dela gehienbat. Laburbilduz, esan daiteke, euskara testuinguru berrietara moldatzeko gai dela, betiere hiztunen komunitatearen egunerokotasuneko errealitatearekin modu estuan lotuta. Honek erakusten digu, globalizatutako eta etengabe konektatutako mundu honetan ere, euskaldunek badutela gaitasun berezi bat beren lekua bilatu eta bertan finkatzeko.

Bibliografia

Agerri, R. & Rigau, G. (2016). Robust multilingual Named Entity Recognition with shallow semi-supervised features. In *Artificial Intelligence*, 238, 63-82.

Al Zamal, F.; Liu, W. & Ruths, D. (2012). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *ICWSM*, 270, 2012.

Bastian, M.; Heymann, S. & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *ICWSM*, 8 (2009), 361-362.

Blei, D. M.; Ng, A. Y. & Jordan, M. I. (2003). *Latent dirichlet allocation*. *Journal of machine Learning research*, 3(Jan), 993-1022.

Binkley, D.; Heinz, D.; Lawrie, D. & Overfelt, J. (2014). Understanding LDA in source code analysis. In *Proceedings of the 22nd international conference on program comprehension* (pp. 26-36). ACM.

Blondel, V. D.; Guillaume, J. L.; Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. In *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

Eckert, P. (2017). Age as a sociolinguistic variable. In *The handbook of sociolinguistics*, 151-167.

Hong, L. & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80-88). ACM.

Marquardt, J.; Farnadi, G.; Vasudevan, G.; Moens, M. F.; Davalos, S.; Teredesai, A. & De Cock, M. (2014, January). Age and gender identification in social media. In *Proceedings of CLEF 2014 Evaluation Labs* (pp. 1129-1136).

Morgan-Lopez, A. A.; Kim, A. E.; Chew, R. F. & Ruddle, P. (2017). Predicting age groups of Twitter users based on language and metadata features. In *PloS one*, 12(8), e0183537.

Nguyen, D.; Gravel, R.; Trieschnigg, D. & Meder, T. (2013, July). How Old Do You Think I Am? A Study of Language and Age in Twitter. In *ICWSM*.

Nguyen, D.; Do ruöz, A. S.; Rosé, C. P. & De Jong, F. (2016). Computational sociolinguistics: A survey. In *Computational linguistics*, 42(3), 537-593.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O. & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. In *Journal of machine learning research*, 12(Oct), 2825-2830.

Rao, D.; Yarowsky, D.; Shreevats, A.; & Gupta, M. (2010, October). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (pp. 37-44). ACM.

Rehurek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Rosenthal, S. & Mckeown, K. (2011, June). 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 (pp. 763-772). Association for Computational Linguistics.

Sievert, C. & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Steyvers, M. & Griffiths, T. (2007). Probabilistic Topic Models in *Latent Semantic Analysis: A Road to Meaning*, Landauer, T. and Mc Namara, D. and Dennis, S. and Kintsch, W., eds.

Umap.eus (2018, May 30). Rankinga: Orokorra. Retrieved from https://umap.eus.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E. P.; Yan, H. & Li, X. (2011, April). Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval* (pp. 338-349). Springer, Berlin, Heidelberg.

Oharrak

1. Ekarpen hau poster moduan aurkeztua izan zen Gasteizko kongresu saioan.

2. Lan hau Eusko Ikaskuntzak diruz lagundutako izen bereko Master Amaierako Lanean oinarrituta dago. 2017-2018 ikasturtean emandako laguntza, 'Gazteak-euskara binomio dinamikoa: erabilera sustatzeko gakoak' egitasmoan txertatzen da.