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

Egilea: Joseba Fernandez de Landa Agirre
Tutoreak: Iñaki Alegria eta Rodrigo Agerri

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko
titulua lortzeko bukaerako proiektua
- - - -
Eusko Ikaskuntzak diruz lagundutako Master Amaierako Lana

2018ko urria. Donostia, Euskal Herria

Saila: Lengoaia eta Sistema Informatikoak.

Laburpena

Teknologia berrien etengabeko garapenak aldaketak eragin ditu gizakion arteko komunikazio moduetan. Honela, geroz eta ohikoagoa da sare sozialak eguneroko bizitzan erabiltzea, inolako mugarik gabeko komunikazioa ahalbidetuz. Komunikazio-esparru birtual honek hartueman publiko hauek jasotzeko aukera ematen du, Twitterren adibidez, testuan oinarritutako informazio mordoa edukiko da eskuragarri. Komunikazio-esparru berri honen sorrerak eta berau ustiatzeko aukerak ikerketa esparru berri bat irekitzeko aukera ematen du, ikerketa-teknika berriak beharko dituena. Aukera berri honi esker, euskararen etorkizunarekin erlazionatutako ikerketa burutzea izango da asmoa, hizkuntza honen erabilera Twitterren aztertuz. Ikerketa honetan arreta berezia jarriko da pertsona gazteetan, esparru berri hauetan nagusi izateaz gain, etorkizuna baitira. Horretarako euskal txiolariengan zentratuko da ikerketa, hauek bi taldeetan zatituz: gazte eta heldu. Behin banaketa burututa, talde bakoitzak ze gairi buruz hitz egiten duen eta zeinekin harremantzen diren azaleratzea izango da asmoa. Datu mordo hauek kudeatzeko lengoaia naturalaren prozesamenduko (NLP) teknikak erabiliko dira, ikerketa sozialerako konputazio zientzien teknikak aplikatuz.

Abstract

The continuous advance of new technologies, has generated changes in the way of relating between humans. Therefore, it is increasingly common to use social networks in our day to day, allowing communication without limits. This new virtual space of communication allows to collect the ways of relating, on Twitter for example, we will have access to a lot of information based on text. The creation of this new communication space and the possibility of mining it, allows the creation of a new field of research that needs new research techniques. Thanks to this new opportunity, the intention will be to carry out an investigation related to the future of basque language, analyzing the use of this language on Twitter. In this research special attention will be placed on young people, since they are the majority in these new spaces as well as being the future. For this the research will focus on Twitter users who speak basque, separating these into two groups: youth and adults. Once the separation is made, we will inquire about the topics they talk about or how the relationships in each group are given. To manage this massive data, natural language processing (NLP) techniques will be used, applying computer science techniques to social research.

Gaien aurkibidea

1 Sarrera

Azken urteetan geroz eta ohikoagoak bilakatzen ari dira sare sozialak, gizakion bizitzaren zati geroz eta handiagoa hartzen ari baitira. Gure artean erlazionatzeko bide berriak ireki dira honi esker, oztopo espazial zein denboralak hautsi egin dira eta etengabeko konexioa ahalbidetu da komunitatearekin, nonahi eta noiznahi komunikatuta egoteko aukera irekiz (Castells et al., 2005). Edonorekin, edozein momentutan, edozein tokitik konektatuta egoteko aukerak, harremantzeko modu hauen arrakastaren atzean daude, gazteen artean batez ere, eguneroko bizitzan normaltasunez txertatuz. Harremana oinarri duten kanal birtual hauek, ikerketarako aproposak diren informazio iturri berriak kontsideratu ditzakegu, gizarte ikerkuntzari ikerketa esparru berriak gehituz.

Abagune berri honen aurrean, Eusko Ikaskuntza elkarteak, euskaraz hitz egiten duten gazteak nola harremantzen diren eta zertaz aritzen diren ezagutzeko asmoarekin, proiektu hau finantzatu du. Eusko Ikaskuntzak diruz lagundutako Master Amaierako Lan honetan, gazte euskaldunengana heltzeko modurik onena teknologia berrien bitartez dela erabaki da, aipatutako informazio iturri berriak irekiz eta ikerkuntza teknika aurreratuak erabiliz. Ikerketa honen metodo zein teknika berritzaileak aplikatzeko, beharrezkoa izan da IXA taldearen inplikazio zuzena, batez ere, masterraren bitartez trebetasunak eta ezagutza transmititzeko. Honekin, ikus daiteke Euskal Herriko testuinguruan aukerak badaudela erronka berriei aurre egiteko eta begi puntuan dauden gaiei lotzeko.

Sare sozialetan euskal hiztun gazteek dauzkaten harreman sareen eta jorratzen dituzten gaien analisia egitea izan dira lan honen helburuak. 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 Hizkuntzaren prozesamenduko tekniketan oinarritu gara, informatikako teknologiak gizarte-ikerkuntzan aplikatuz. Teknologia berrietan sortzen den edukia teknologia berrietan oinarritutako ikerketa-tekniken bitartez aztertuak izango dira, esparru berriak ikerketa-teknika berrien bidez ikertuz. Asmo horrekin, Twitter sare soziala hautatu da, identifikatutako euskal komunitate bat daukan sarea delako, eta baita, informazioa erauzterako orduan erraztasunak ematen dituelako ere, orokorrean jarioa publikoa delako, Instagram, Facebook eta Snapchat sareetan ez bezala.

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 iradokitzeko saiakera eginez. Adinaren identifikazio automatikoaren bideragarritasun falta ikusita, hurbilpen bat egitea erabaki da, txioaren idazteko eran oinarritu gara erabiltzaile gazteak eta helduak ezberdintzeko hizkuntza ez-

formalaren erabilera maiztasunean oinarrituz. Ezberdintze hau burutzea garrantzitsua da, gazteen errealitatea ezagutzea lehentasun bat baita lan honetan, bai gazteen eduki eta harremanak ezagutzeko, eta baita, ezkutuan dauden ezaugarri demografikoak antzemateko kapazitatea badagoela frogatzeko ere. Gazteak eta helduak sailkatzeko hurbilpena burututa egonda, bi izango dira ikerketa honetan argitu nahiko diren inkognitak, euskaldunek zertaz hitz egiten duten eta norekin harremantzen diren. Alde batetik, euskal txiolariek zein gairen inguruan hitz egiten duten ezagutzea izango da lehendabiziko asmoa, komunitate honen gairik errepikatuenak zeintzuk diren argituz. Honetarako, Hizkuntzaren prozesamenduko teknikak (NLP) erabiliko dira, testuetatik informazioa ateratzeko. Beste aldetik, euskal txiolariak nola harremantzen diren ezagutzea izango da bigarren asmoa, egindako birtxioetan oinarrituta, komunitateak zehaztu eta hauetako pertsona garrantzitsuak identifikatuz.

Gizarte-zientzien eta konputazio-zientzien arteko konbinazioan kokatzen da lan honen ekarpen nagusia, aurrerapen teknologikoetan eta informazio mugagabean oinarritutako gizarte likidoa (Bauman, 2015) interpretatzen eta ulertzen lagunduko duen sinbiosia. Interneteko sareari esker, gizartea hobeto ulertzeko teknikak garatzeko orduan mugarri izan nahi du lan honek, gizarte-ikerkuntzan ireki den bide berri honetan lehenengo pausoak emanez. Lehenik eta behin, bilketan iraultza txiki bat piztea lortu da, interneteko datu-iturriak ustiatzeari esker, datu-iturri mugagabe bati ateak irekiz. Gizarte zientzietan datuak lortzea gutxi batzuen esku egon da beti suposatzen duen kostu altuagatik, baina Twitterren adibidez, Instagramen edo Facebooken ez bezala, edozeinek eskuratu ditzake datuak, informazioaren demokratizazio bat emanez. Hala ere, unibertsoa sare sozial zehatz honen erabiltzaileetara mugatzen da, iturri masiboagoak irekitzearen beharra dagoela azpimarratuz. Bigarrenik, datuen prozesaketan ere, hainbat teknika berritzaile aplikatzearen abantailak ikusi ahal izan dira. Zehatzago esanda, sare sozialetako erabiltzaile ezberdinen idatzizko adierazpen oinarrituta, zertaz hitz egiten duten eta nola harremandu diren antzematea lortu da. Honez gain, Ikasketa Automatikoan oinarritutako teknikei esker, ezaugarri demografikoak iradokitzeko ahalmena ere garatu da. Hau da, erabiltzaileen nolakotasunean oinarrituta, adina bezalako ezaugarri demografikoak intuitzeko kapaza den sistema bat garatu da. Honi esker, gizarte ikerlariontzat oinarrizkoak diren datuak (demografikoak kasu) igartzeko kapazitatea badagoela erakutsi da, datuak eskuragarri ez edukitzearen arazoa saihestuz. Hirugarrenik eta azkenik, datuen interpretagarritasuna errazteko asmoarekin, erabilitako teknikek bistaratze intuitibo eta grafiko bat izango dute, irudien laguntza izanik.

2 Proiektuaren definizioa

2.1 Ikergaia

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, hauen atzean dagoen harremanen-sarea islatuz.

2.2 Ikerketaren helburuak eta hipotesiak

Helburuak

Ikerketa lan honen helburu nagusia, **Twitter sare sozialera konektatuta dauden euskal hiztun gazteen on-line errealitatea ezagutzea** izango da. Helburu nagusi horren lorpenerako, tarteko 4 helburu finkatu dira:

a) **Ikerketa soziala ahalbidetuko duen datu-iturri berriak irekitzea**.

b) **Gazteak eta helduak identifikatzea, txio informalen kontzentrazioan oinarrituta**. Sailkatzaile bat sortuko da, euskarazko txio formal eta informalak desberdintzen dituena. Idazkeran oinarrituz, ezaugarri demografikoak iradokitzeko hurbilpena egitea.

c) **Gazteen eta helduen gaiak zeintzuk diren identifikatu eta konparatzea**. Euskal txiolariek publikatutako idazkietan oinarrituta, gai ohikoenak iradokitzea izango da asmoa.

d) **Gazteen eta helduen harremanak zeintzuk diren identifikatu eta konparatzea**. Euskal txiolarien eduki trukean oinarrituz, hauen arteko harremanak azaleratzea izango da asmoa.

Hipotesiak

a) Sare sozialetatik euskal erabiltzaileen informazioa lortu daiteke modu merke batean.

b) Gazteek zein helduen ezberdintasuna txioak idazteko moduan oinarritu daiteke, txio informalen kontzentrazioan oinarrituz.

c) Gazteak eta helduak zein gairi buruz aritzen diren identifikatu daiteke txioen testuan oinarrituta.

d) Gazteen eta helduen harremanak zeintzuk diren antzeman ahalko da birtxioetan oinarrituta.

2.3 Metodologia

Euskal txiolari gazteen errealitatea aztertuko duen ikerketa lan hau burutzeko, lana hainbat zatitan burutua izan da. Lehenik eta behin datuen erauzketa gauzatu beharko da, Twitter sare sozialetik beharrezkoak diren datuak batuz. Bigarren pausua, gazteak identifikatuko dituen sailkatzaile bat garatzea izango da, txioen testuaren nolakotasunean oinarrituta, gazteak eta helduak ezberdinduz. Behin Twitterreko euskal erabiltzaileen multzoa gazte eta heldu artean banatuta egonda, hauen gaiak eta harremanak zeintzuk diren argitzera igaroko da. Euskal txiolarien gaiak identifikatzeko, erabiltzaile bakoitzaren euskarazko txio pertsonaletan oinarrituko gara, bertatik gaiak ondorioztatzeko. Euskal txiolarien harremanak ezagutzeko, euskarazko birtxioetan oinarrituko gara, zeinek zein birtxiokatu duen oinarri izanda, 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-en 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 (Roesslein, 2009) paketea erabili da, Twitterreko APIa deitzen duen paketea. Pakete honi esker, euskal txiolarien zerrendako erabiltzaile guztien informazioa erauzi ahalko da, kontutan hartuta APIak mugak dituela (gehienez 3200 txio erabiltzaileko eta 15 minuturo gehienez 15 erabiltzaile erauzi daitezke).

- **Gazte/heldu identifikatzailea:** Atal honen helburua, euskal txiolari gazteen identifikazioa izango da, txiolariak gazte eta heldu artean ezberdinduko direlarik. Sailkapen hau, ez da adinaren arabera egingo, mezuen edukiaren 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. Helburu honetarako, corpus txiki bat etiketatu da, 1.000 txio eskuz etiketatuz informal

eta formal moduan. Txioak formal eta informal artean bereiziko dituen sailkatzailea entrenatzeko 3 metodo ezberdin erabili dira, ez-gaibegiratutako eredu estatistikoa, ikasketa automatikoa eta *IXA pipes*en dokumentu sailkatzailea:

Ez-gaibegiratutako eredu estatistikoari dagokionean, txio bakoitza euskarazko *hizkuntz eredu* formal batekin konparatu da, formala edo informala den zehazteko. Txioa eta *hizkuntz eredu* formalaren distantzia neurgarri bihurtzeko asmoarekin *perplexity* neurria erabiliko da, geroz eta altuagoa izanik informalagoa dela adieraziz. Horrela, *hizkuntz ereduarekiko distantzia*ren kontzeptua erabiliko da, hizkuntza barietate bat beste batetik zein ezberdina den adierazten lagunduko duena.Txio formal eta informalak bereizte aldera, txioak euskarazko *hizkuntz eredu* formal batekin konparatuak izango dira eta hauen *hizkuntz ereduarekiko distantzia* kalkulatu da. hizkuntz ereduen arteko distantzia hau kalkulatze aldera, karaktereetan oinarritutako n-grama modeloan oinarritu gara, 7-gramak erabiliz, ereduen arteko distantzia kalkulatzeko asmoarekin (Gamallo et al., 2017). Hizkuntz ereduarekiko distantzia kuantifikatzeko *perplexity* balioa erabili ohi da (Chen et al., 1999), beraz, ebaluatutako txio bakoitza hizkuntz eredu formaletik zenbat aldentzen den adieraziko digun balioa lortzeko *perplexity*a erabiliko da.

Ikasketa Automatikoari erreparatuz, python programako *scikit* metodoa (Pedregosa et al., 2011) erabili da hau aplikatzeko. Etiketatutako corpusaren tamaina txikia kontutan hartuta, entrenamendua eta ebaluazioa *cross-validation* bitartez egitea erabaki da 5 iterazio erabiliz. Metodo honetan, hainbat sailkatzaile ezberdin frogatuko dira, besteak beste: Logistic Regression, Naive Bayes, k-NN, Decision Tree, Random Forest eta SVM.

*IXA pipes*en dokumentu sailkatzailea (Agerri eta Rigau, 2016), ikasketa automatikoan oinarrituta egongo da, baina beste datu-iturri batzuk lagungarri edukiko ditu ere. Sistema honek informazio lokala konbinatzen du etiketatu gabeko testu kantitate handietan induzitutako ezaugarrien clusterrekin (Agerri eta Rigau, 2016). Era honetan, etiketatutako corpus txikiaren datu falta orekatu egiten da, etiketatu gabeko testutik erauzitako ezaugarrien clusterrei esker. Hala ere, etiketatutako corpusaren tamaina txikia kontutan hartuta, entrenamendua eta ebaluazioa *cross-validation* bitartez egin da 5 iterazio erabiliz.

- **Txiolarien ohiko gaiak identifikatu:** Datu moduan euskarazko txio pertsonalak erabili dira gaien identifikaziorako. Horretarako, Topic-Modeling izeneko teknika erabiliz eta LDA algoritmoa aplikatuz. Zehazki LDA aplikatzeko gensim paketea (Rehurek eta Sojka, 2010) erabili da, era honetan, LDA aplikatu ahalko da lortutako testuaren baitan. Bistaratze intuitibo bat lortzeko, emaitzak irudi moduan argitaratzen dituen *LDAvis* metodoa (Sievert eta Shirley, 2014) erabili da, interpretazioan lagungarria izango dena. Zehazki *pyLDAvis* paketea erabili da, *LDAvis* metodoa pythonera egokituz (Sievert eta Shirley, 2014).

- **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, erabiltzaileak nola harremantzen diren oinarri hartuta, azpitaldeak ezagutzeko.

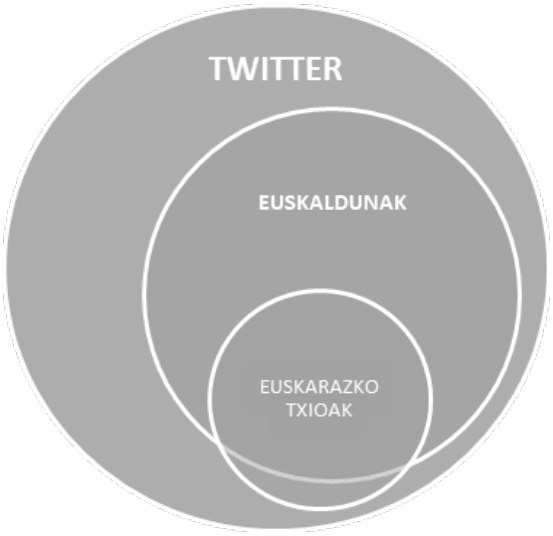
3 Datuen erauzketa

Euskal txiolari gazteen gai eta harremanak zeintzuk diren identifikatzeko asmoarekin, lehenengo pausua datuen bilketa izan da. Datu hauek lortzeko, lehenik eta behin aztertu beharreko unibertsoa mugatu dugu. Hau da, ikertu behar ditugun subjektuak identifikatu dira, geure kasuan, euskaraz egiten duten txiolariak identifikatuz. Gero, erabiltzaile edo txiolari hauen txioak erauzi ditugu, kontutan hartuta Twitterrek honetarako API bat daukala. Azkenik datu hauen azaleko azterketa edo ebaluazioa egin beharko da, jakiteko datu hauek egokiak diren. Besteak beste, erauzitako euskarazko txioen kopurua ezagutzeko asmoarekin.

3.1 Euskal txiolarien identifikazioa

Esan bezala, lehenik eta behin, hasierako pauso moduan euskal txiolariak identifikatzean oinarritu gara, Twitterren euskara erabiltzen duten erabiltzaileak hain zuzen. Twitterreko sare erraldoian euskarazko erabiltzaileak topatzeari ekin zaio, horretarako, euskaraz txio kopuru minimo bat argitaratzen duten erabiltzaileen identifikazioa burutuz. Honez gain, erabiltzaile bakoitzaren txio kopuru minimo bat euskaraz izatea beharrezkotzat jo da, erabiltzaile ez euskaldunek ere euskarazko txioak eduki ditzaketelako, edukia partekatu baitaiteke Twitterren. Horrela, euskarazko txioak argitaratu dituzten subjektuetatik, euskal erabiltzaileak hautemateko, euskarazko txio kopuru minimo bat duten erabiltzaileak hautatuko dira. Era honetan, erabiltzaile euskaldunak direla baieztatu daiteke zehaztasun gehiagorekin, nahiz eta kopuru txiki bat kanpo geratuko den.

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. Honela, geure corpusa osatzen hasteko lehenengo eginbeharra beteta egongo litzateke, erabiltzaile euskaldunen zerrenda esanguratsu bat lortzearena.



Irudia 1: Unibertsoaren identifikazioa.

3.2 Twitterretik datuak erauzi

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. Datu-bilketa hau burutu ahal izateko hainbat modu ezberdin daude, baina hiru nagusi ezberdinduko dira hurrengo lerroetan eta bakoitzaren ezaugarrien arabera geure atazarako erabiliko dena aukeratu da, ustiatu nahi ditugun datuetan oinarrituz;

Streaming bidezko erauzketa: metodoari termino zerrenda bat pasatzeko aukera dago eta metodo honek momentu horretatik aurrera txiokatutako elementuak itzuliko dizkigu, termino zerrenda horietako bat topatzen baldin badu.

Termino bidezko erauzketa: termino konkretu bat pasatzen diogu funtzioari eta 15 minuturo 45.000 txio inguru lortzeko aukera dago. Lortutako txio bakoitzean termino hori azalduko da, lortutako emaitzak termino horrekiko menpekoak izango direlarik. Bestalde, metodo honekin soilik azkeneko asteko txioak lortu daitezke, txio zaharrak lortzeko aukera zailduz eta erauzketa mugatuz.

Erabiltzaileen erauzketa: aurreko bi puntuetan bilaketa termino konkretuen arabera burutzen da, honetan erabiltzaile zerrenda batean oinarrituta, erabiltzaile bakoitzak publikatutako txioak erauzten dira. Erabiltzaile bakoitzeko 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.

Gure ikerketaren berezitasunak ikusita eta erabiltzaileen zerrenda luze bat daukagula kontutan hartuta, **erabiltzaileen erauzketa** egitea aukera egokiena izan da. Honek esan nahi du zerrendako erabiltzaile bakoitzaren txioak erauziak izango direla, erabiltzaile bakoitzetik azkeneko 3.200 txioak lortuz gehienez jota.

Behin erauzketa teknika definituta egonda, erauzketarekin hasi gara, 2018ko maiatzaren 30 eta 31an. Orotara 7980 euskal txiolarien azkeneko txioak erauztea lortu da, guztira 10 milioi txio baino gehiago batuz.

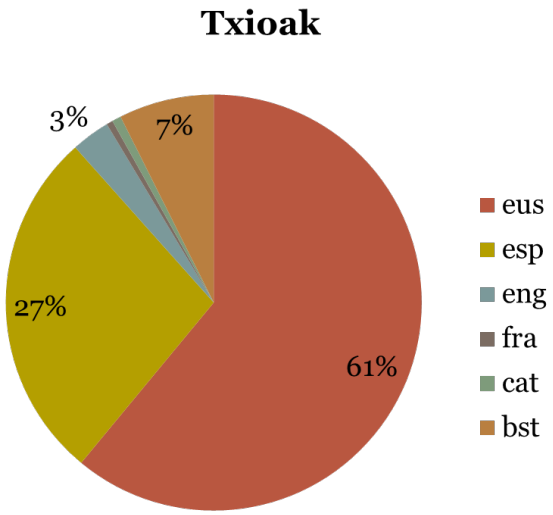
3.3 Datuen azaleko azterketa

Esan bezala, erabiltzaileetan oinarritutako erauzketa burutu ostean, 10 milioi txio baino gehiago lortu dira. Lan honen helburua euskarazko edukiak zeintzuk diren identifikatzean oinarritzen denez, txio guzti hauek hizkuntzaren arabera sailkatu behar dira. Hizkuntzaren araberako sailkapen hau egiteko, txioen meta-datuetara joko dugu, txio bakoitzak idatzia izan den hizkuntzaren etiketa daramalako, hortaz ataza erraza izan da txioak hizkuntzaren arabera sailkatzea.

Txioak hizkuntzaren arabera nola banatzen diren ikusi aurretik, bi multzo nagusi bereizi dira, txio pertsonalak eta birtxioak banatuz. Bi multzo hauek egitea erabaki da, bakoitzak sortutako edukia partekatzen denarekin ezberdintzeko asmoarekin. Multzoketa burutu ostean, txio pertsonalek datu bolumen osoaren %49 osatuko lukete, 5 milioi txio baino gehiago izanik. Birtxioek ostera, geratzen den %51 osatuko lukete, 5 milioi baino gehiagokoa ere. Behin txio eta birtxioak talde banatuetan sailkatuta daudela, hizkuntzaren araberako analisiari ekingo diogu.

Txio pertsonalak (5.198.043): Birtxioak alde batera utzita, errepikatu gabeko txioetan zentratuko gara. Kasu honetan, ia bi heren (%61) euskaraz daudela egiaztatu da. Gazteleraren presentzia ere indartsua da (%27) euskal txiolarien testu sorkuntzan, ehuneko esanguratsu bat lortuz. Ingelesa (%3) eta frantsesaren (%0,5) emaitza baxuek erakusten dute gazteleraren nagusitasuna euskaldunen bigarren hizkuntza moduan. Bestalde, %7a azaltzen zaigu *beste hizkuntzak* atalean, baina atal honen zati handiena web-orrietarako estekek edo erabiltzaile izenek osatzen dute.

Gazteen identifikaziorako eta erabiltzaileen gaiak zeintzuk diren identifikatzeko beraz, txioen euskarazko %61 horrekin geratu beharko gara. Guztira euskarazko 3.171.485 txio pertsonal izango dira erabilgarriak ataza hauetarako, corpusaren zati txiki bat dela ematen duen arren, kopuru handi bat dela argi utziz.

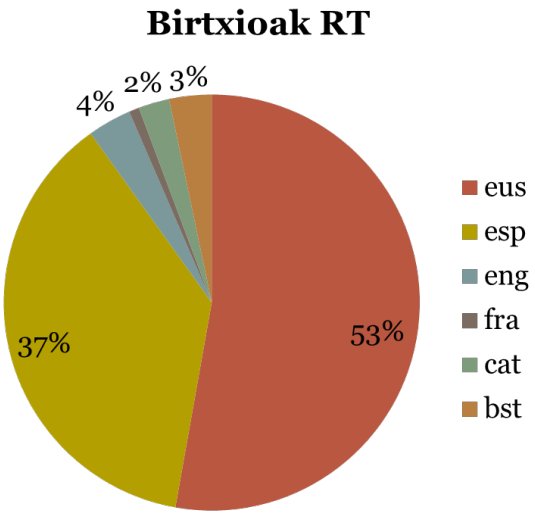


Irudia 2: Txio pertsonalen hizkuntza.

Birtxioak (5.473.031): Lehenik eta behin aipatu beharra dago birtxioen kopurua txio pertsonalena baino handiagoa dela, hala ere uste baino txikiagoa. Hau da, bilaketak termino bidez egiten direnean, txioen trafikoaren %80 inguru birtxioak dira, baina kasu honetan nahiko orekatua dagoela esan daiteke. Bestalde, topic modelinga burutzerako orduan datu hauek kanpoan geratuko dira, aipatu bezala, errepikatutako informazioa ekiditeko asmoarekin.

Hizkuntzaren banaketari erreparatuz, deigarria da nola euskarazko txioen ehunekoa nabarmen jaisten den birtxioen kasuan, %61a izatetik %53ra igaroz. Horrez gain, azpimarratu beharra dago ere gaztelerazko txioen hazkuntza, %27 izatetik, %37 izatera igaro direlako. Honek erakusten digu txiolari euskaldunek ohitura handia daukatela gaztelerazko edukiak partekatzeko.

Erabiltzaileen harremanak euskaraz nola ematen diren identifikatzeko, beraz, birtxioen euskarazko %53 honekin geratu beharko gara. Guztira euskarazko 2.891.136 birtxio erabiliko dira harreman euskaldunak nola ematen diren ikusteko, soilik euskarazko birtxioak kontutan hartuz.



Irudia 3: Birtxioen hizkuntza.

Hizkuntzaren inguruko gogoeta moduan esan daiteke, euskal txiolariek askotan gaztelera erabiltzen dutela edukia sortzeko. Horrez gain, edukia partekatzerako orduan, gaztelera are gehiago erabiltzen dela ikusi daiteke. Ondorio nagusia genekiena da: gaztelerak pisu handia daukala euskal txiolarien komunitatean, baina lan honetan kuantifikatzea lortu da, hauen komunikazioaren ehuneko handi bat hizkuntza honetan gertatzen dela frogatuz.

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

Taula 1: Erauzitako datu erabilgarriak, euskarazko txioen nolakotasuna.

4 Gazte/heldu identifikatzailea

Identifikatzaile hau, lanaren bigarren pausua izango da; behin datuak erauzita, gazteak zeintzuk diren identifikatzeari ekin zaio. Behin euskal txiolarien datuak lortu direla, hauetatik gazteak zein helduak zeintzuk diren iragartzea izango da asmoa, horretarako txio informalen kontzentrazioan oinarrituz, gerora, bi talde hauetan gaiak eta harremanak nolakoak diren konparatzeko asmoarekin.

Atal hau euskal txiolari gazteak identifikatzeko sistema bat garatzean oinarrituko da, gazteak eta helduak bereizten dituen identifikatzaile bat sortuz. Horretarako metodologia ezberdinak erabili dira, metodoen artean konparaketa bat egiteko asmoarekin, haien artean, ez-gainbegiratutako metodo estatistikoa, ikasketa automatikoan (*Machine Learning*) oinarritutako hainbat sailkatzaile, zein *IXA pipes* dokumentu sailkatzailea erabili dira. Horrela, hiru sistemetatik emaitza onenak lortzen dituena zein den aztertuko da eta zergatia argitu.

Euskal txiolarien artean gazteak eta helduak ezberdinduko dituen sistema hau inplementatzeko, lehenik eta behin bibliografian esandakoa jaso da. Behin atal honekin erlazionatutako lanak jasota eta aztertuta, kasu zehatz honetarako metodo egokiena aukeratu da, hau da, sailkatzailearen mota zein izango den aukeratu eta hau iragartzeko erabiliko diren ezaugarriak zehaztuko dira. Klaseak zein ezaugarriak definituta daudelarik, sailkatzailea garatzea izango da hurrengo pausua, hiru modelo ezberdin erabilita: eredu estatistikoan oinarritutako sailkatzaile ez-gainbegiratua, ikasketa automatikoan oinarritutako hainbat sailkatzaile eta *IXA pipes*en dokumentu sailkatzailea. Sailkatzaile ezberdinak garatu ostean, emaitza onenak lortzen dituena aukeratu eta datu errealetan aplikatua izango da, gure corpusa osatzen duten erabiltzaileen artean helduak eta gazteak zeintzuk diren antzemanez.

4.1 Erlazionatutako lanak

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 sexua izanik ohikoenak (Cesare et al., 2017). Kasu zehatz honetan, Twitterreko erabiltzaile euskaldunen bizitza etapa identifikatzea izango da asmoa, erabiltzaile hauek gazte edo helduen artean ezberdinduz. Gazte/heldu identifikatzaile honekin, bi bizitza etapen arteko konparaketa egitea izango da intentzioa, gai eta harremanetan ezberdintasunak antzemateko.

Twitterren adinaren detekziorako, artearen egoerako metodo aipagarrienak bildu dira jarraian azaltzen zaigun taulan (2. taula). Taula horretan metodo bakoitzaren erreferentzia, sailkatzaile mota eta asmatze tasa ikusi daitezke. Taula honetan ikusi daitekeen moduan, sailkatzaile ohikoenak ataza honetarako *Logistic Regression* (Nguyen et al., 2014; Morgan-Lopez et al., 2017) eta *Support Vector Machine* (Rao et al., 2010; Al Zamal et al., 2012; Marquart et al., 2014) dira. Sailkatzaile guzti horiek ikasketa automatikoan oinarritutako metodo gainbegiratuak dira, guztiak tamaina esanguratsu bateko etiketatutako corpus batean oinarrituta. Ikerketa guzti horietan datuen etiketatzea erabiltzaile

bakoitzaren adinean oinarrituta dago, eskuz erabiltzaile kopuru esanguratsu bat anotatu dutelarik.

Sistema	**Sailkatzailea**	**Asmatze tasa**
Rao et al. (2010)	*Support Vector Machine*	% 74,11
Al Zamal et al. (2012)	*Support Vector Machine*	% 80,50
Marquart et al. (2014)	*Support Vector Machine*	% 48,31
Nguyen et al. (2014)	*Logistic Regression*	% 86,32
Morgan-Lopez et al. (2017)	*Logistic Regression*	% 74,00

Taula 2: Bibliografiako sistemak, adinaren detekzioa Twitterren.

Jarraian sistema bakoitzean murgilketa txiki bat egingo da, etiketatutako corpusa eta klase kopurua ezagutzeko asmoarekin (3. taula). Sistemak oso ezberdinak dira ikusi daitekeen moduan, behar ezberdinetara egokitutako ikerketak baitira. Alde batetik, eskuz etiketatutako corpus hauen tamaina ikerketaren arabera aldatuz doa, baina guztiek 300 eta 3000 erabiltzaile artean dauzkate etiketatuta. Bestalde, bibliografiako sistema onenek, beren sailkapena burutzeko asmoarekin, adin tarte bitarrak (Rao et al., 2010; Al Zamal et al., 2012) edo hirutarrak (Nguyen et al., 2014; Morgan-Lopez et al., 2017) erabili dituzte, geroz eta klase gutxiago edukiz (bitarra edo hirutarra) sailkatzeko ataza errazago burutzen dela frogatuz.

Erabiltzaile bakoitza adinaren arabera sailkatzeko asmoz, lortutako txio bakoitzeko testua zein meta-datuak bezalako aldagaiak erabili dira. Honela, artearen egoerako autore ezberdinek, bakoitzak bere ezaugarri propioak aukeratu ditu, txioen testuaz gain, erabiltzaile zehatzen meta-datuetan oinarritutako hainbat ezaugarri ere konbinatuz. Hala ere, sistema hauek bata bestearengandik oso ezberdinak direla ikusi den arren, guztiek testua hartzen dute kontutan, gehienak sailkatzailea txioen idazkera estiloan oinarritzen direlarik (Rao et al., 2010; Al Zamal et al., 2012; Nguyen et al., 2014; Morgan-Lopez et al., 2017).

Sistema	**Corpus tamaina**	**Klase kopurua**	**Hizkuntza**
Rao et al.(2010)	1000 erabiltzaile	2	Ingelera
Al Zamal et al.(2012)	400 erabiltzaile	2	Ingelera
Marquart et al.(2014)	306 erabiltzaile	5	Gaztelera
Nguyen et al.(2014)	3110 erabiltzaile	3	Nederlandera
Morgan-Lopez et al.(2017)	3184 erabiltzaile	3	Ingelera

Taula 3: Bibliografiako sistemak, adinaren detekzioa Twitterren.

Amaitzeko, esan beharra dago sailkatzaile bakoitzaren asmatze tasari (*accuracy*) erreparatuz gero, oraindik bide luze bat geratzen dela adina bezalako ezaugarri demografikoak egoki iragartzeko sare sozialetan. Adinaren sailkatzaileen asmatze ahalmena perfekziotik urruti daudela kontutan hartuta, ikerketa soziala zaildu egiten da, oinarrizkoak diren datu

demografikoetan errore altuak daudelako. Adina bezalako datuk gizarte ikerketaren oinarrietako bat dira eta errore altuak emaitzak desitxuratu ditzake. Hala ere, ezin dugu ahaztu, txikia baldin bada ere, errorea beti egongo dela, gizartea konplexutasunez beteta baitago eta orokortze perfektu bat izatea ezinezkoa izango litzateke. Horregatik, errorea izanda ere, gizarte multzo zehatz honen orokortasunak jasotzeko arazorik ez da egongo, euskal txiolari gazteen eta helduen arteko konparaketa orokor bat egitea baita asmoa.

4.2 Metodoa

Artearen egoerako sistema ezberdinak ikusi ostean, antzeman daiteke sistema onenak bitarrak edo hirutarrak direla, sistema sinpleenekin emaitza onenak lortzen direla erakutsiz. Honez gain, aztertutako sistema guztiek komunean daukate, adin tarte konkretuen arabera sailkatzen dutela. 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, 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.

4.2.1 Txio formal eta informalen sailkatzailea

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, formalak eta informalak. Sailkatailea beraz, txio informal zein formalak sailkatzean oinarrituko da, erabiltzaile bakoitzaren txio guztiak honen arabera sailkatuz. Horrela, txio pertsonal bakoitzaren testuan oinarrituko da gure sailkatzailea, testua nola idatzia izan den aztertuz, formal edo informal moduan etiketatuko duena. 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 itxura 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.

Behin euskarazko txioen testua aldagai moduan aukeratuta, testu hau garbitu eta estandarizatzeari ekingo zaio, erroreak ekidin eta sailkapena hobeto burutzeko. Era honetan, karaktere alfanumerikoak dituzten hitzak soilik mantenduko dira, emotikono, erabiltzaile izen (@), hastag (#) edo url esteka guztiak ezabatuz eta testua sinplifikatuz. Honez gain, behin karaktere alfanumerikoak soilik dituzten hitzak dauzkagula, bakarrik lau hitz baino gehiago dituzten txioak hartuko dira kontutan, beste txio guztiak alboratuz.

Erabiltzaileen bizitza etapa zein den iragartzen lagungarri izango den formal/informal txio sailkatzailea garatzea izango da hurrengo pausua. Txio bakoitzaren idazkera estiloa ezberdintzen lagunduko duen sistema garatzeari ekingo zaio beraz. Helburu hori betetzeko, ez-gainbegiratutako modelo estatistiko zein ikasketa automatikoko metodoak erabiliko dira, azkeneko hauek artearen egoeran emaitza onak eman baitituzte. Horietaz gain ere, IXA taldearen baitan garatutako *IXA pipes*en dokumentu sailkatzailea erabiliko da, etiketatutako corpus txikietarako aukera ona delarik. Hala ere, sailkatzaile ezberdinak garatzera igaro aurretik, hauek garatu eta ebaluatzeko erabiliko den corpus txiki bat etiketatzeari ekin zaio.

4.2.2 Corpusaren etiketatzea

Kontutan hartuta erreferentziazko aldagaia erabiltzaile bakoitzaren txio pertsonalak direla, corpus guztiaren zati txiki bat etiketatu da, sailkatzaile ezberdinak garatu eta ebaluatzeko asmoarekin. Horretarako, ausaz 1.000 txio pertsonal aukeratu dira eta idazteko moduaren arabera sailkatuak izan dira. Era honetan, txio batek *formal* etiketa eramango du, hizkuntza estandarrean idatzia izan bada, edo *informal* etiketa, txioa modu kolokialean idatzia izan denean. Corpusaren egiturari begiratuz gero (4. taula), ikusi daiteke 1.000 txioez osatzen den etiketatutako corpus honek txio formal eta informalen kopuru antzekoa daukala, corpus orekatu bat izanik. Txioen batez besteko luzera ia 10 tokenekoa da, txio laburrenak 5 tokenekoak izanik eta luzeenak 34 tokenekoak izanik.

Etiketatutako corpusaren ebaluazioa	
Txio kopurua	1.000
Txio formalak	492
Txio informalak	508
Txio laburrena	5 token
Txio luzeena	34 token
Tokenak txioko batezbeste	9,66

Taula 4: Eskuz etiketatutako corpusaren ebaluazioa.

4.3 Ez-gainbegiratutako eredu estatistikoa, *perplexity*

Ez-gainbegiratutako eredu estatistikoan oinarritutako metodo hau garatzeko, txioak euskarazko *hizkuntz eredu* formal batekin konparatuko dira, horiek formalak edo informalak diren zehazteko. Eredu ezberdin horien distantzia neurgarri bihurtzeko asmoarekin *perplexity* neurria erabiliko da, geroz eta altuagoa izanik informalagoa dela adieraziz.

Horrela, atal honetan, *hizkuntz distantzia*ren kontzeptua erabiliko da, hizkuntza barietate bat beste batetik zein ezberdina den adierazten lagunduko duena.Txio formal eta informalak bereizte aldera, txioak euskarazko *hizkuntz eredu* formal batekin konparatuak izango dira eta hauen *hizkuntz ereduarekiko distantzia* kalkulatu da. Hizkuntz ereduen arteko distantzia hau kalkulatze aldera, karaktereetan oinarritutako n-grama modeloan oinarritu gara, 7-gramak erabiliz, ereduen arteko distantzia kalkulatzeko asmoarekin (Gamallo et al., 2017). Hizkuntz ereduarekiko distantzia kuantifikatzeko *perplexity* balioa erabili ohi da (Chen et al., 1999), beraz lan honetan ere ebaluatutako txio bakoitza hizkuntz eredu formaletik zenbat aldentzen den adieraziko digun balioa lortzeko *perplexity*a erabiliko da. Gainera, ez da lehenengo aldia *perplexity*aren balioa txioetan osagai informalak aurkitzeko erabiltzen dela, hiztegi kanpoko hitzak (OOV) detektatzeko erabilia izan baita txioak oinarri hartuta, txio informalak hautemateko (Gonzalez, 2015). Kasu horretan hitzak modu soltean aztertuak izan ziren arren, kasu zehatz honetarako txioa osorik hartu da kontutan, orokortasun bat bilatzeko asmoarekin.

4.3.1 Diseinua

Txio bakoitza, karaktereen 7-grametan oinarrituz, hizkuntz eredu formalarekin konparatu ostean, txioaren urruntasuna kuantifikatzeari ekingo zaio, txio bakoitzari *perplexity* balio bat luzatuz. *Perplexity*aren balioak adierazten du, modelo bat zeinen ondo egokitzen den ebaluazio datuetara, modelora egokitzeko probabilitatearen alderantzizkoa izanik. Honela, txioak geroz eta balio altuagoa lortu, geroz eta urrunago egongo da hizkuntz eredu formaletik, hau da, geroz eta balio altuagoak esango du informalagoa dela txio hori. *perplexity* balioa txikiagoa baldin bada ostera, txioa hizkuntz eredutik hurbil dagoela adieraziko du, txioa formala dela adieraziz. Beraz, txio formal eta informalak bereizteko *perplexity* balio muga bat jarri beharko da, atalase hori baino baxuagoak formal moduan kontsideratuz eta

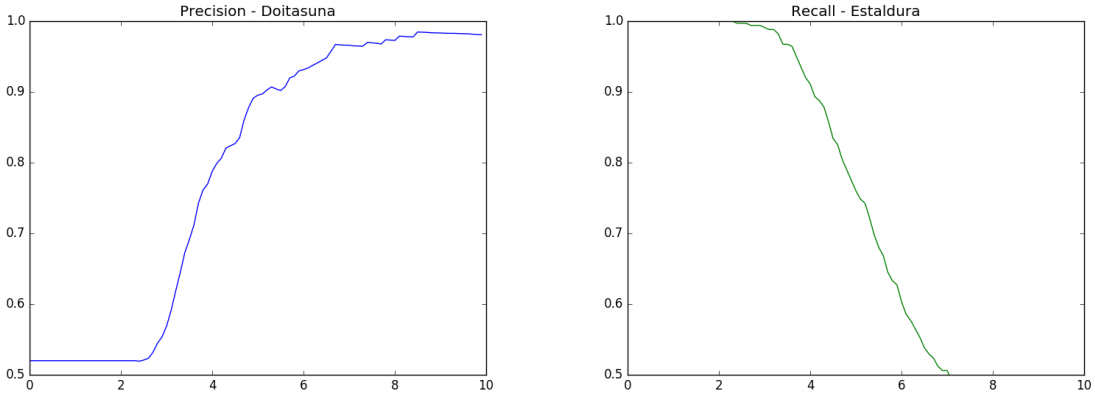
hortik gorakoak informaltzat joz.

7-grametan oinarritutako eredu hau garatzeko, *dataset* edo corpusa bi zatitan banatu da, alde batetik garapen faserako corpusaren %65a erabili da eta bestetik, ebaluazio faserako corpusaren %35a. Garapen fasean informal/formal bereizteko *perplexity*ak hartu beharko lukeen balioa zehaztuko da, ahalik eta txio gehien ondo sailkatzeko intentzioarekin. Ebaluazio fasean ostera, garapen fasean finkatutako atalaseak lortutako emaitzak ikusiko dira, garapenean erabili ez diren datuak baliatuz. *Dataset*a bi zatitan banatuta, garapena eta ebaluazioa, sistemak datuen gehiegizko egokitzea jasatea ekidingo da, emaitzak aurrez ikusi gabeko datu multzo batetan balioztatuko baitira.

4.3.2 Ereduaren garapena: atalasearen moldaketa

Txioak sailkatzeko sistema garatuta dagoenean, txio bat formaltzat jotzeko *perplexity*ak hartu beharko lukeen balio minimo egokia aukeratu beharko da. Azaldu den moduan, *perplexity*aren balioa geroz eta handiagoa izan, txioa aurrez zehaztutako *Language Model*etik geroz eta gehiago aldentzearen seinale izango da. Hortaz, balio zehatz bat eman beharko zaio informal eta formalak ezberdintzen dituen balioari. Hau da, atalase bat finkatuko da, non balio horretatik behera txioak formalak direla kontsideratuko da, eta balio horretatik aurrera, txioak informaltzat joko dira.

Sistemaren garapenerako etiketatutako korpusaren %65a erabiliko da, hau da, *perplexity*aren balioa zein izango den zehazteko txio bat informala den edo ez erabakitzeko. Atalasea markatuko duen balio hau, 0tik 10era doazen balio guztiekin probatu da, 0.01 iteratuz aldiro. Honela, jarraian datozen grafikoetan (4, 5 eta 6 irudiak), atalasearen muga balioa moldatu ahala, doitasuna, estaldura, asmatze tasa eta F-scorea nola aldatzen diren ikusi ahalko da. Honek lagunduko digu, *perplexity*aren balio finko bat zehaztean txio informal zein formalak ondo sailkatuko dituena.



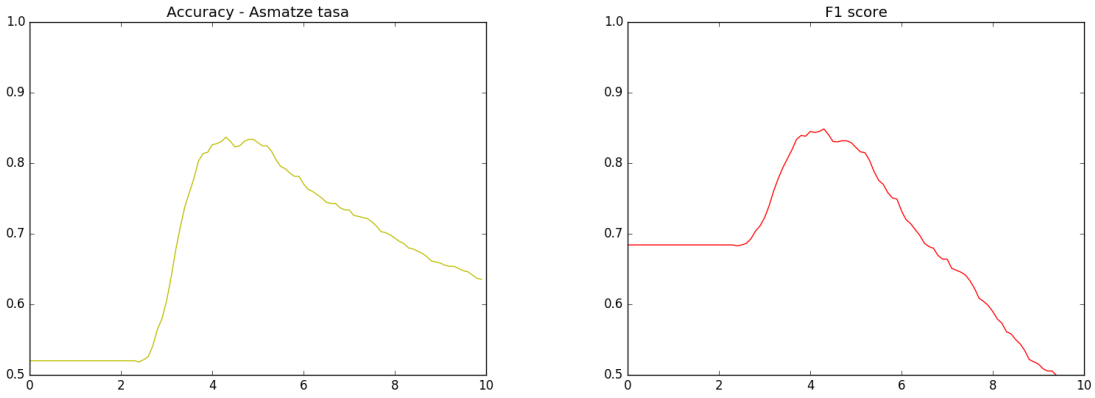
(a) Doitasuna (b) Estaldura

Irudia 4: Doitasuna eta estaldura.

Doitasunari esker, sistemak txio informalak sailkatzeko duen gaitasuna ikusi dezakegu. Honek esan nahi du, ikusi ahalko dela sistemak txio informalak ondo sailkatzeko duen gaitasuna, formal/informal muga balioa moldatzen goazen heinean. Horrela, 4a irudian ikusi daitekeen moduan, *perplexity*ak 3 balioa baino altuagoa lortzen duenean, *baseline*a (0.5) gainditzen hasten da, honek esan nahi du, muga 3 baino balio altuagoetan jartzean, hasten dela sailkapen hobea egiten. Bestalde, *perplexity*aren muga 7 edo balio altuagoetan zehaztean, hobekuntza gelditu egiten da, txio informal ia guztiak ondo sailkatzen dituelarik. Hala ere, doitasuna balioarentzat 0.8ko balio bat desiragarria izanik, atalasea *perplexity*=4 baliotik gora kokatu beharko genuke gutxienez. Baina, doitasun eta estaldura desiragarri batzuen artean oreka bat lortu beharko da.

Bestalde, **estaldura**ri erreparatzen bazaio, sistemak txio informalak detektatzeko duen gaitasuna ikusi daiteke, muga balio edo atalasea moldatzen den heinean. Honela, 4b irudian ikusi ahal da, *perplexity*ak 3 balioa baino altuagoa lortzen duenean, estaldura galtzen hasten dela. Honek esan nahi du, 3 baliotik aurrera txio informalak okerrago detektatzen hasten dela. Beste behin ere, onargarritzat hartuko dugu 0.8ko estaldura izatea gutxienez, horretarako atalasearen balioa 4.5 baino txikiagoa izan beharko delarik. Hala ere, atalasearen balio finkoa aukeratzeko aurretik lortutako doitasun desiragarri baten emaitzak ere kontutan hartu beharko dira.

Hortaz, bi neurri ezberdin hauen arteko oreka bat bilatu beharko da atalasearen balioa finkatu eta emaitza onak lortu nahi izatekotan. Batetik sistemak txio informalak ondo sailkatzea (doitasuna) interesatzen zaigu, eta bestetik, txio informal horiek identifikatzea (estaldura) interesatzen zaigu. Grafikoei erreparatuz gero (4. irudia), ikusi daiteke balio bat hobetuz doan heinean bestea okertu egiten dela. Honek esan nahi du, geroz eta txio informal gutxiago identifikatu ahala, sailkapen lan hori hobeto egingo duela. Horregatik, bi lerroen ebaketa puntuari arreta berezia jarri beharko zaio, eta baita, bi balioak konbinatzen dituen F-score neurriari ere. Jakinda doitasun desiragarria 4tik aurrera dela eta estaldura desiragarria 4.5etik behera, esan daiteke jarraibide batzuk baditugula atalasearen balio egokia aukeratzeko.

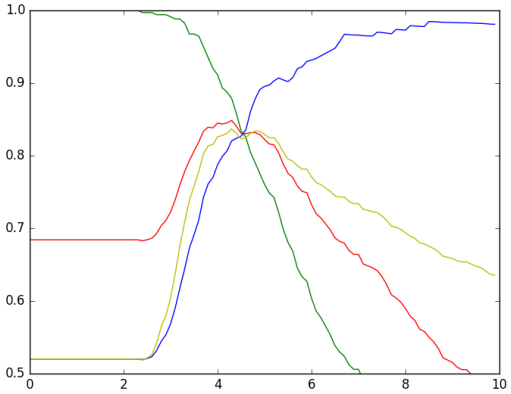


(a) Asmatze tasa (b) F-neurria

Irudia 5: Asmatze tasa eta f-neurria.

Sistemaren egokitasuna aztertzeko beste neurri bat, **asmatze tasa** dugu, sistemaren egokitasun orokorra azalduko duena. Orain arte estaldura eta doitasunarekin txio informalak detektatzeko gaitasuna ebaluatzen ibili gara, baina neurri honekin txio informal zein formalak sailkatzeko gaitasuna ebaluatuko dugu. Hortaz, asmatze tasarekin, txioak formal edo informalak diren ondo sailkatzeko kapazitatea neurtuko da. Atalasearen balioa mugitu ahala, ikusi daiteke balioa 4 ingurura iritsi ahala emaitza onenak lortzen dituela, 5a irudian ikusi daitekeen moduan. Beraz, txioak *Language Model*arekiko *perplexity* balioaren muga 4 inguruan kokatu beharko da sailkapen egoki bat egiteko eta ahalik eta errore gutxien lortzeko.

Behin doitasuna eta estaldura balioak aztertu direla, bi balio hauek konbinatzen dituen **f-neurria** aztertzeari ekingo zaio. 5b irudian ikusi daitekeen moduan, 4 balio inguruan emaitza onenak lortzen direla ikusi daiteke. Asmatze tasaren balio onenekin gertatu den moduan, 4 balio inguruan aurkitzen diren balio horietan arreta jarriko da erabakia hartzerako orduan. Atalaseak 4 edo inguruko balioak hartzen dituenean, 0.8tik gorako *f-neurria*k lortzen ditu, honi gehituta estaldura eta doitasun balio desiragarrienak 4 - 4.5 tartean daudela, erabakia hartzeko datu nahikoak daudela ondorioztatu daiteke.



Irudia 6: Aurreko neurri guztiak batera.

Sistemaren garapenean lortutako emaitzen arabera, **atalaseari 4.4 balioa ematea** erabaki da, horrela txio bakoitzak *Language Model*ekiko lortutako *perplexity* balio honetatik gora txioak informalak direla kontsideratuko da, eta balio honetatik behera, formalak direla kontsideratuko dira. Atalaseari balio hori egokitzea erabaki da, asmatze tasa altuenetarikoa lortzen delako puntu honetan, 0.831eko balioa lortuz. Txio informalak iradokitzeko orduan, F-neurri altuenetarikoa lortzen da garapeneko puntu honetan, 0.841 balio inguruan kokatuz. Gehitu beharra dago ere, puntu zehatz honetan estaldura (0.858), doitasuna baino zertxobait hobea dela (0.824), aukeratutako balioa bi lerroen arteko ebaketa puntua baino lehenago hartzea erabaki baita (6. irudia eta 5. taula).

Etiketa	**Errorea**	**Totala**	**Doitasuna**	**Estaldura**	**F-neurria**
Informala	48	338	0.824	0.858	0.841
Formala	62	312	0.839	0.801	0.820

Taula 5: Entrenamenduaren balioak, ez-gainbegiratutako eredu estatistikoa.

4.3.3 Ereduaren ebaluazioa

Garapenaren atalean atalaseari balioa egokitu eta gero, aukeratutako eredua ebaluatzeari ekingo zaio. Aurretik azaldu den moduan, atalaseari luzatutako balioa 4.4ekoa izango da. Honek esan nahi du, *Language Model*ekiko 4.4 *perplexity* balioa baino gehiago lortzen duten txioak informaltzat joko direla.

Ebaluazio lan honetarako, garapenerako erabili ez den corpus etiketatuaren zatia erabili da, hau da, geratzen den %35a. Jarraian ikusi daitezkeen tauletan (7 eta 8 taulak), asmatze tasa, doitasuna, estaldura zein *f-neurria* balioak daude, atalasearen 4.4 *perplexity* balioarentzat. Ikusi daitekeen moduan, garatutako sistema estatistiko honek, 0.825ko *f-neurria*a lortzen du txio informalak sailkatzerako orduan, eta baita, %83,57ko asmatze tasa txioak formal edo informalak diren sailkatzean.

	Informal iragarriak	**Formal** iragarriak
Informal errealak	**144**	26
Formal errealak	35	**145**

Taula 6: Konfusio-matrizea, ez-gainbegiratutako eredu estatistikoa.

Testa (atalasea = 4.4)	
Asmatze tasa	**0.826**

Taula 7: Testaren balioak, ez-gainbegiratutako eredu estatistikoa.

Etiketa	**Errorea**	**Totala**	**Doitasuna**	**Estaldura**	**F-neurria**
Informala	26	170	0.805	0.847	0.825
Formala	35	180	0.848	0.806	0.826

Taula 8: Testaren balioak, ez-gainbegiratutako eredu estatistikoa.

4.4 Ikasketa automatikoan oinarritutako eredua

Ez-gainbegiratutako sailkatzailea garatu eta ebaluatu ostean, *Machine Learning* eredua oinarri duten sailkatzaile ezberdinekin sailkatzeari ekingo zaio, artearen egoerako sistema guztiak ikasketa automatikoan oinarritzen baitira.

Sailkatzailea inplementatzeko lehenengo pausu moduan, corpusean zehaztu diren klaseak ondo esleituko dituen corpuseko datu edo ezaugarri egokiak aukeratu beharko dira. Artearen egoerako sistema ezberdinek ezaugarri zerrenda luze eta konplexuak barnebiltzen dituzten arren, geure kasuan, soilik testua aintzat hartuko da. Testuan oinarrituta dagoen sailkatzaile hau garatzeko, txio bakoitza kontutan hartuko da, bakoitzarekin hitzen bektore bat sortuz. Era honetan, sistemaren *input*a 0 zenbakiz gainezka dagoen hitzen bektore erraldoi batean oinarrituko da. Testuan oinarritutako *input*a *Bag of Words* (BoW) eredu baten moduan tratatua izango da, hau da, ordenarik gabeko hitzen multzo baten moduan.

Sistemaren garapenerako erabiliko diren datuen multzoaren tamaina txikia dela eta (1000 txio), erabaki da entrenamendua eta ebaluazioa *cross-validation* bitartez egitea. Era honetan, 5 karpetetan oinarritutako *cross-validation*a burutuko dugu, entrenamendu eta testerako datu guztiak batera erabiliko ditugularik. Honela, geure sistemak *overfitting*a egitea ekidingo dugu, nahiz eta *dataset* edo corpusaren tamaina txikia izan.

Machine Learning Classifier (BoW)	**Asmatze tasa (accuracy)**
5-NN (k-NN)	0.614
Decision Tree	0.677
Random Forest	0.707
Naive Bayes	0.765
Logistic Regression	0.775
SVM (Linear Kernel)	0.777

Taula 9: Ebaluazioaren balioak, *Machine Learning* eredua (5 CV).

Ikasketa automatikoaren teknika ezberdinetan oinarritutako 6 sailkatzaile garatu ostean, lortutako emaitzen azterketari ekingo zaio. 9. taulako emaitzei begiratuta, sailkatzaile onenak *Logistic Regression* eta *SVM* direla ikusi daiteke. Beraz, lortutako emaitzarik onenek zerikusia daukate artearen egoerako beste sailkatzaileekin, artearen egoerako sistemen sailkatzaile berdinak izan baitira emaitza onenak lortu dituztenak. Artearen egoerako sistemekiko ezberdintasun ugari eduki arren, etiketatzeko modua edo ezaugarrien trataera ezberdina kasu, sailkatzaile onenak berdinak direla ikusi daiteke. Hala ere, emaitza baxuak lortu direla esan daiteke, etiketatutako corpusaren tamaina txikiak, ikasketa edo entrenamendurako datu oso gutxi eskaintzen baititu.

4.5 *IXA pipes* dokumentu sailkatzailearen eredua

Ez-gainbegiratutako metodo estatistikoa eta ikasketa automatikoan oinarritutako metodoak jorratu ostean, *IXA pipes* dokumentu sailkatzailea erabiliko da sailkapen atazarako. Txioak informal eta formal moduan sailkatzeko erabiliko den metodo hau, ikasketa automatikoan oinarritua egongo da, baina beste datu-iturri batzuk lagungarri edukiko ditu ere. Sistema honek informazio lokala konbinatzen du etiketatu gabeko testu kantitate handietan induzitutako ezaugarrien clusterrekin (Agerri eta Rigau, 2016). Era honetan, etiketatutako corpus txikiaren datu falta orekatu egiten da, etiketatu gabeko testutik erauzitako ezaugarrien clusterrei esker. Metodo honi esker, eskuz etiketatu beharreko datu kopuru handiekiko dependentzia arindu egiten da (Agerri eta Rigau, 2016), etiketatutako corpus txikiekin sailkapen egoki bat egitea ahalbidetuz. Horrez gain, aipatu beharra dago, sailkatzerako orduan abiadura handiz egiten duela, beste ereduak baino bizkorrago eginez. Sailkatu beharreko datu multzoaren tamaina handia dela kontutan hartuta, bizkortasunak ere garrantzia izango du beste ereduekiko konparaketan.

Sistema honek eredu gainbegiratuak ikasten ditu, horretarako Perzeptroiaren algoritmoa (Collins, 2002) erabiltzen duelarik. Sailkapena, etiketatutako corpusarekin eta etiketatu gabeko testutik erauzitako ezaugarrien clusterren konbinaketari esker burutzen da. Era honetan, informazio lokala, hitzen irudikapenen (clusteringa) ezaugarrien hiru mota ezberdinekin konbinatuko da: Brown clusterrak (Brown et al., 1992), Clark clusterrak (Clark, 2013) eta Word2Vec clusterrak (Mikolov et al., 2013). Sailkatzaile honek gainera, euskarazko hitzen irudikapenen clusterrak garatzeko aukera ematen du ere, lan honetan burutu beharreko atazarako aproposa izanik. Hitzen irudikapenak burutzeko aipatuta-

ko hiru teknika ezberdinetatik (Brown, Clark eta W2v) eratorritako clusterrak pilatu eta konbinatzen dira, emaitza hobeak lortuz (Agerri eta Rigau, 2016). Clusterren ezaugarri horiek, token bakoitzari talde batekiko kidetasuna ematen diote, horrela ikusi gabeko tokenak, ikusitakoekin erlazionatzen dira cluster berdinean azalduz gero (Agerri eta Rigau, 2018).

4.5.1 Ereduaren garapena: hitzen irudikapenen cluster kopurua aukeratu

Garapena *cross-validation* bitartez egin da, etiketatutako datuen multzoaren tamaina txikia dela eta (1.000 txio). Kasu honetan, 5 *fold*etan oinarritutako *cross-validation*a burutu da, sistemak *overfitting*a egitea ekidinez, nahiz eta etiketatutako corpusaren tamaina txikia izan. Sistema honen berezitasuna informazio lokala, hitzen irudikapenen (clusteringa) ezaugarrien hiru mota ezberdinekin konbinatzean datza. Horrela, Brown, Clark eta W2V clusterren hainbat konbinaketa ezberdin frogatzeari ekin zaio, Elhuyar web corpuesan eta Tokikomen corpusean, teknika bakoitzetik cluster kopuru egokiena hautatuz corpus bakoitzerako. Elhuyar corpuserako, Brown teknikaren bidez hainbat cluster kopuru frogatu dira, arrakastatsuena 3200 klaserekin izanik. Elhuyarreko corpusarekin jarraituz, Clark teknikaren bidez 600 klase izan dira egokienak eta W2V teknikarako 300 klase izan dira emaitza onenak lortu dituenak. Tokikomeko corpusean operazio berdina errepikatu da, Brownerako 0, Clarkerako 600 eta W2Verako 300 klase aukeratuz. Entrenamenduko teknika honi buruzko argipen gehiago topatu daitezke Agerri eta Rigauen (2016, 2018) lanetan.

Honela, hitzen irudikapenen ezaugarrien konbinaketa ezberdinak frogatu ostean, Elhuyar web corpuserako: Brown 3200 klase, Clark 600 klase eta Word2Vec 300 klase; eta Tokikomeko corpuserako: Clark 300 klase eta Word2Vec 500 klase; konbinaketekin osatutako eredua aukeratu da, emaitza onenak lortu baititu. Honela, etiketatutako 1.000 txioekin 5 *fold*etako *cross validation* egin ostean % 88,5eko asmatze tasa lortu da. Bestalde, txio informalak detektatzeko kapazitateari begira doitasuna (0.889) estaldura (0.884) baino apur bat altuagoa dela ikusi daiteke, nahiz eta ezberdintasuna minimoa da. Bestalde, txio informalak detektatzerako orduan 0.886ko f-neurria dauka, txio formalak ondo detektatzerako orduan f-neurria 0.883koa denean. Ikusi daiteke txio informalak detektatzeko orduan txio formalak detektatzeko baino zertxobait hobeto egiten duela. Honela, emaitza altuenak lortu direla kontuan hartuta, hautatutako ezaugarriekin ebaluazioa egin beharko da.

Etiketa	**Errorea**	**Totala**	**Doitasuna**	**Estaldura**	**F-neurria**
Informala	59	508	0.889	0.884	0.886
Formala	56	492	0.881	0.886	0.883

Taula 10: Entrenamenduaren balioak, *IXA pipes* dokumentu sailkatzailea.

4.5.2 Ereduaren ebaluazioa

Behin garapen fasea amaituta dagoela, aukeratutako ezaugarrien araberako (Elhuyar web corpusa: Brown 3200, Clark 600 eta W2V 300; Tokikom: Clark 300 eta W2V 500) eredu

bat entrenatu eta ebaluatu da. Ebaluazioa burutzeko, aukeratutako ezaugarriak oinarri hartuta, eredu bat entrenatuko dugu horretarako etiketatutako 650 txio erabiliz. Behin eredua entrenatuta dagoela, ebaluazioari ekingo zaio, horretarako etiketatutako corpusetik entrenamendurako hartu ez diren 350 txioekin testa burutuz. Era honetan, sailkatzailearen fidagarritasuna frogatu ahalko da, esleitutako atazaren egokitasuna azaleraziz.

Testaren emaitzei erreparatzen badiegu, lehenik eta behin aipatu beharra dago Asmatze tasa % 86,57koa dela (12. taula), nahiko emaitza altua dela, batez ere kontutan hartzen bada etiketatutako corpusaren tamaina txikia dela. Hala ere, argi ikusten da hobekuntza tarte handi bat geratzen dela oraindik, etiketatutako corpusa oso txikia baita. Doitasunari (*Precision*) begira, 0.864 balioarekin, ikusi daiteke informaltzat jo diren txio batzuk formalak direla, sistemak gain-sorkuntza txiki bat duela erakutsiz, errore txikia izan arren, emaitza honek erakusten du badagoela lana emaitzak hobetzeko. Estaldurari (*Recall*) erreparatzen badiogu, 0.859 balioarekin, ikusi daiteke sistema kapaza dela txio informalak nahiko ondo detektatzeko, ia gehienak barneratu arren, atal honetan ere emaitzak hobetu daitezkeela erakutsi da. Nahiz eta, doitasunaren balioa estaldurarena baino apur bat handiagoa izan, esan daiteke bi neurrien arteko aldea txikia dela eta nahiko orekatuta daudela. Amaitzeko, F-neurriari begira, 0.861ekoa dela ikusi daiteke, estaldura eta doitasuna neurrien balioen artean, bi neurri hauen arteko konbinaketa baita.

	Informal iragarriak	**Formal** iragarriak
Informal errealak	**146**	24
Formal errealak	23	**157**

Taula 11: Konfusio-matrizea, *IXA pipes* dokuementu sailkatzailea.

Testa	
Asmatze tasa	**0.866**

Taula 12: Testaren balioak, *IXA pipes* dokuementu sailkatzailea.

Etiketa	**Errorea**	**Totala**	**Doitasuna**	**Estaldura**	**F-neurria**
Informala	24	170	0.864	0.859	0.861
Formala	23	180	0.867	0.872	0.870

Taula 13: Testaren balioak, *IXA pipes* dokumentu sailkatzailea.

4.6 Emaitzen analisia

Behin modelo ezberdinetan oinarritutako sailkatzaileak garatu ostean, guztien artean onena aukeratzeari ekin zaio, arrakastaren zergatiak argituz. Garapen edo entrenamendurako *dataset*aren tamaina txikia kontutan hartu beharko da ere ebaluazioa burutzerakoan, sailkatzaileek izan ditzaketen datu-iturri lagungarrietan (*metodo ez-gainbegiratu* eta *IXA pipes*

kasu) arreta berezia jarriko delarik analisi honetan.

Lehenbizi, esan beharra dago ikasketa automatikoan oinarritutako ereduek zailtasunak eduki dituztela datu hauek sailkatzeko, entrenamendurako corpusa txikiegia delako. Datu kopuru txiki honek sorrarazten du, sistemak soilik berezitasunak ikasiko dituela, ebaluazio orduan asmatze tasa desegoki bati bide emanez gain-sorkuntzaren erruz. Arazo hau konpontzeko asmoarekin, etiketatutako corpusa handitzea izango litzateke konponbideetako bat, orokorragoak diren datuak ikasiz. Hala ere, etiketatzea ataza zail eta nekeza izateaz gain, ikasketa datuekin sortzen diren hitzen bektoreen dispertsioa handia izaten jarraituko luke, hobekuntza nabarmenki zailduz. Era honetan, Ikasketa Automatikoan oinarritutako sailkatzaile onenak Erregresio Logistikoa (*logistic regression*) eta SVM (*Support Vector Machine*), beste metodoen emaitzetatik urrun geratzen dira.

Ez-gainbegiratutako eredu Estatistikoan oinarritutako *Perplexity*aren neurketak ikasketa automatikoan oinarritutako metodoek baino hobeto egin du, SVM sailkatzailearen emaitza asko hobetuz. Hobekuntza nabarmen hau, hizkuntz ereduak eskaintzen duen datu kopuru gehigarrian oinarritua egon daiteke, ikasketa automatikoan oinarritutako sailkatzaileek baino datu gehiago baitauzka garapen faserako. Hortaz, emaitza hobeak lortu ditu hizkuntz ereduak datu kopuru ugari eskaintzen baititu. Datu multzo gehigarri hori lagungarri izanda ere, metodo estatistikoetan oinarritutako sistema hau ez da onena izan.

Baieztatu da *IXA pipes* dokumentu sailkatzailea sistema onena dela Twitterren euskal txiolari gazteak identifikatzeko, txio informaletan oinarrituz. Ikasketa automatikoan oinarrituta egon arren eta etiketatutako corpusa txikia izanda ere (1000 txio), bestelako datuak lortzeko aukerari esker (hitzen irudikapena), emaitza altuak lortu dira, garatu diren beste sistemen emaitzak nabarmen gainditurz. Hitzen esanahi semantikoa jasotzen duten clusterrek gehitu dioten informazioari esker lortu ditu emaitza onenak, hobekuntza garrantzitsua lortzen dela agerian utziz.

Argi ikusten da sailkatzaile estatistiko ez-gainbegiratua eta *IXA pipes* dokumentu sailkatzailea hoberen egiten dutenak direla. Hala ere, bien artean konparaketa egitea beharrezkoa da, horretarako ebaluazio faseko emaitzak alderatuko direlarik. Bi sailkatzaile hauek konparatu ahal izateko, train eta test set berdinak erabili dira, datu multzo berdinekin sistema bakoitza frogatuz gerora konparatu ahal izateko. Lehenik eta behin, asmatze tasari begira, ikusi daiteke *IXA pipes* sailkatzaileak *Perplexity* neurrian oinarritutako sailkatzailea baino lau puntu hobeto egiten duela (14. taula). Modu berean, f-neurriari erreparatzen bazaio, *IXA pipes*en emaitzak askoz hobeak direla ikusi daiteke (15. taula). Emaitzak aztertu ostean, ikusi daiteke *IXA pipes* dokumentu sailkatzailea metodo onena dela txio informal eta formalak sailkatzerako orduan, bestelako sistemek baino emaitza aski altuagoak lortzen baitira sistema honekin.

Sailkatzailea	**Asmatze tasa**
Perplexity	% 82,57
IXA pipes	% **86,57**

Taula 14: Testaren balioen konparaketa sistema onenekin.

Sailkatzailea	Klasea	Doitasuna	Estaldura	F-neurria
Perplexity	Informala	0.805	0.847	0.825
Perplexity	Formala	0.848	0.806	0.826
IXA pipes	Informala	0.864	0.859	**0.861**
IXA pipes	Formala	0.867	0.872	**0.870**

Taula 15: Testaren balioen konparaketa sistema onenekin, klasearen arabera.

Garatutako sistemak aztertzerako orduan, ikusi da sistemak geroz eta informazio gehiago daukanean, sailkapen hobea egiten duela. Ondorio hau ebidentziaren frogapena besterik ez izan arren, ate berri bat irekitzen du ikertzaile xumeentzat, etiketatutako corpusa txikia izan arren, posible egiten duelako emaitza onak lortzea. Era honetan, etiketatutako corpus txiki bat, etiketatu gabeko datu ugarirekin konbinatzen ditugunean garapenaren fasean, emaitza hobeak lortzen dira. Egun, etiketatu gabeko datuei zentzua edo ordena emateko aukera ugari daude, besteak-beste NLP atazetarako hitzen irudikapena (*word representation*) erabiliz. Honela, sailkapenaren atazak emaitza hobeak lortzen ditu egun sarean eskuragarri dauden egiturarik gabeko datuak lagungarri izanik. Horregatik, bereziki garrantzitsua da *IXA pipes*en dokumentu sailkatzaileak eremu honetan egindako ekarpena, hainbat hitzen irudikapen ezberdinak konbinatzen baititu, emaitzak hobetuz.

Garatu den sistema onena ariketa praktiko erreal batean aplikatuko da hurrengo atalean, txiolari gazteak eta helduak desberdintzeko erabiliko dena idazteko eran oinarrituz. Honela, *IXA pipes* dokumentu sailkatzailea erabiliko da, Twitterretik erauzitako corpus erraldoia sailkatzeko asmoekin, txio bakoitza formal eta informal artean banatuz. Erabiltzaileen txio informalen kontzentrazioan oinarrituta, gazteen eta helduen artean ezberdintzeari ekingo zaio, adinaren arabera egin ordez, idazteko moduaren arabera eginez. Sailkapen honek, gerora bereizitako bi talde hauen gai ohikoenak eta harremanak azaleratzeko aukera emango du. Era honetan, modu independentean landu ahalko da talde bakoitza eta lortutako emaitzak alderatuz, heldu eta gazteen ezberdintasun zein puntu komunak zeintzuk diren argitara emango da.

4.7 Sailkatzailea corpusean aplikatuz

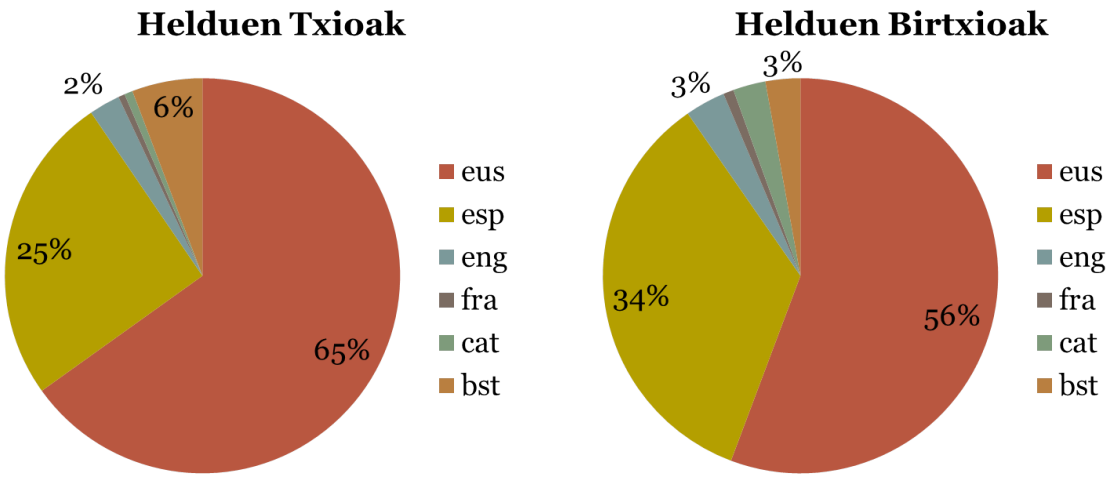
IXA pipes dokumentu sailkatzailea erabilita, erauzitako 7.980 euskal txiolarietatik, 7.087 erabiltzaile sailkatzea lortu da, gutxienez 10 txio euskaraz dauzkaten erabiltzaileak soilik sailkatu direlarik. Sailkatzaileak txioak informal eta formal artean sailkatzen dituenez, txio informalen kontzentrazioan oinarrituko dugu erabiltzaile baten gaztetasuna. Hasiera batean txioen %50a informala izatea ezarri zen erabiltzailea gaztetzat jotzeko. Hala ere, azterketa kualitatibo sinple bat eginda, txio informalen muga %45era jaistea erabaki zen, horrela erabiltzaile gazteen kopuru minimo bat bermatzen delarik. 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 independentean aztertuko denez, tamainaren arabera moldatu beharko dira aplikatuko diren ikerketa-teknika ezberdinak. Lehenik eta behin, heldu eta gazteen multzoak bereizi direla, azaleko azterketa bat burutu da, talde bakoitzaren baitan txioen hizkuntza ezberdinak nola banatzen diren ikusteko asmoarekin.

- **Helduak**

Helduei dagokien datu multzoari erreparatuz, 8 milioi txio baino gehiagoz osatutako corpus handi xamarra daukagu tamainari begira. Sakonago aztertuz, birtxioak(4.345.500) eta txio pertsonalak (4.046.512) orekatuta daudela esan behar da, nahiz eta birtxio kopurua apur bat altuagoa izan. Birtxioen kopurua altuagoa izateak erakusten du joera bat dagoela helduen artean edukia konpartitzerako orduan, sorkuntza propioa baino, partekatzea, ohikoagoa izanik. Hizkuntzei erreparatuz gero, ikusi daiteke euskara dela hizkuntza nagusia bai txio pertsonaletan eta baita birtxioetan ere. Bigarren hizkuntza aipatuena gaztelera dugu, txioen laurden bat baino gehiago hizkuntza honetan ematen direlarik. Bestelako hizkuntzek, paper oso txikia betetzen dute euskal txiolari helduen sorkuntza edo trukean. Amaitzeko azpimarratu beharra dago, edukia konpartitzerako orduan, euskararen erabilera jaitsi egiten dela txio pertsonalekin konparatuz gero, beti ere, gaztelerak hartzen duelarik euskarak usten duen esparru hori.

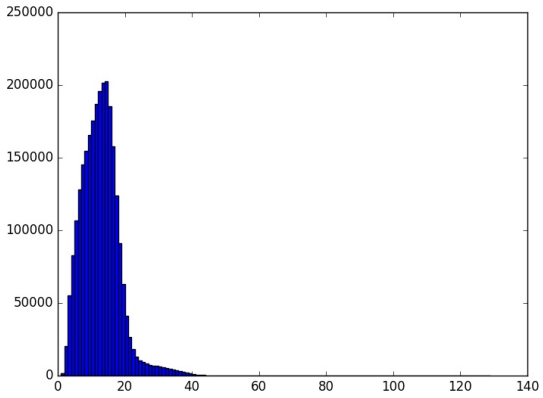


(a) Helduen Txio pertsonalak (4.046.512). (b) Helduen Birtxioak (4.345.500).

Irudia 7: Helduen corpusa, 5.508 erabiltzaile.

Helduen txioen gaineko hizkuntzaren inguruko gogoeta burutu ostean, euskarazko txio pertsonaletan zentratuko gara. Helduen taldean 2.634.534 euskarazko txio pertsonal publikatu dira, 32.273.119 tokenez osatuta. 8. irudian ikusi dezakegu helduen txio pertsonal hauek nola banatzen diren luzeraren arabera, normalenak 4-17 tokenetako luzera tartean kokatzen direlarik. Txioen luzerari begira, batez beste, 12,25 tokenetako luzera daukate helduen txioek, laburrenak token bakarra daukalarik eta luzeenak 126 token dauzkalarik. Batez besteko horrekiko desbiderapen tipikoa 5,65ekoa dugu, beste txioak batez beste zenbat desbideratzen diren adierazten diguna. Bestalde, mediana 12 tokenetan kokatzen da eta moda 14 tokenetan, 200.000 txio baino gehiago izanik 14 tokenetakoak.

HAP masterra

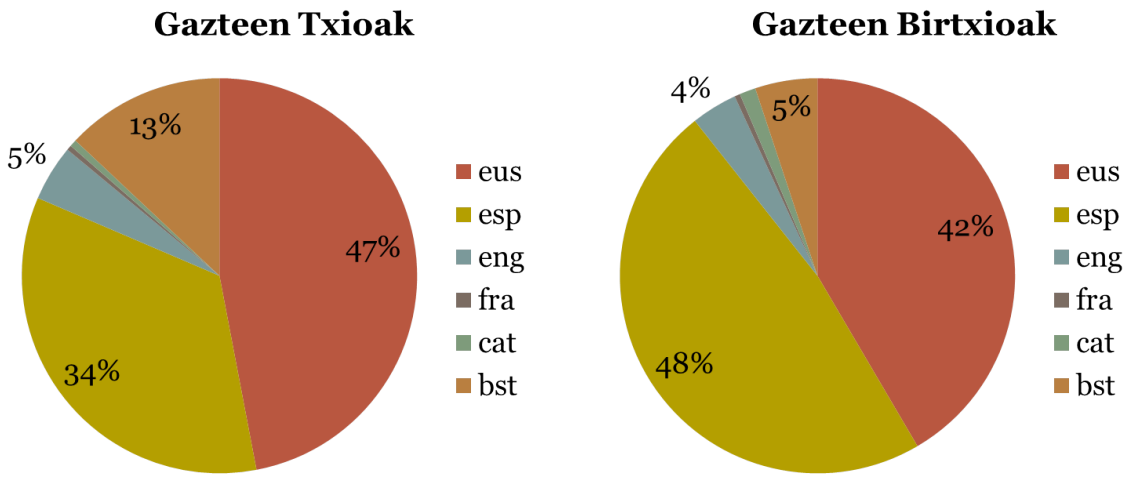


Irudia 8: Helduen Txio pertsonalen tamainaren banaketa, txioen token kopuruaren arabera.

- **Gazteak**

Gazteen atalera jauzi egiten badugu, helduen corpusaren tamainaren laurdena daukan 2 milioiko corpusa daukagu. Sakonera joz, ikusi daiteke birtxio (963.668) kopurua txio pertsonalena (1.128.124) baino zertxobait txikiagoa dela, helduekin gertatzen denaren alderantzizkoa emanez. Honek erakusten du, gazteek edukia partekatzeko ohitura gutxiago daukatela helduekin konparatzen baldin bada, sorkuntza propioari leku gehiago utziz. Bestalde, txio pertsonal zein birtxioen hizkuntzei erreparatzen badiegu, argi ikusten da euskararen presentzia nabarmen txikiagoa dela helduekin konparatuta, txio pertsonaletan helduek baino % 18 gutxiago delarik eta birtxioen kasuan helduek baino % 24 gutxiago izanik. Euskararen presentziaren galera hau, gazteleraren mesedetan izango da berriz ere, txio pertsonalen kasuan heren bat (% 34) lortuz eta birtxioen kasuan ia erdia (% 48) lortuz. Birtxioen kasuan, gazteleraz emandako birtxioak, euskaraz emandakoak baino gehiago direla ikusi daiteke, edukia konpartitzeko ohituran gaztelera gehien erabiltzen dena dela erakutsiz. Honek erakusten du, gazte bezala kontsideratutako erabiltzaile hauek euskara askoz gutxiago erabiltzen dutela helduekin konparatuz gero.

HAP masterra

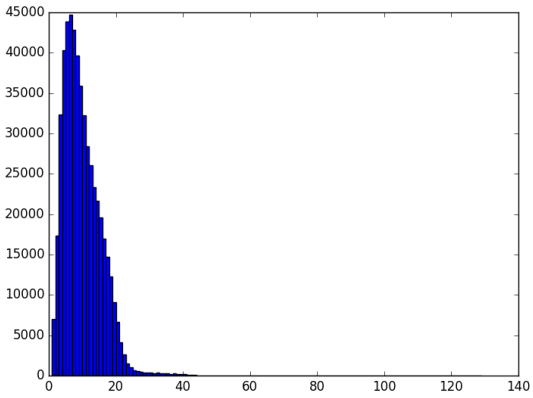


(a) Gazteen Txio pertsonalak (1.128.124). (b) Gazteen Birtxioak (963.668).

Irudia 9: Gazteen corpusa, 1.579 erabiltzaile.

Gazteen, euskarazko txio pertsonaletan zentratuz, 530.226 euskarazko txio pertsonal publikatu dira, 5.018.179 tokenez osatuta. 10. irudian ikusi dezakegu gazteen txio pertsonal hauek nola banatzen diren luzeraren arabera, normalenak 3-13 tokenetako luzera tartean kokatzen direlarik. Txioen luzerari begira, batazbeste, 9,46 tokenetako luzera daukate gazteen txioek, laburrenak token bakarra daukalarik eta luzeenak 62 token dauzkalarik. Batez besteko horrekiko desbiderapen tipikoa 5,49ekoa dugu, beste txioak batez beste zenbat desbideratzen diren adierazten diguna. Helduen taldearekin konparatzen badugu, batez bestekoa 3 token txikiagoa da gazteen txioetan, nahiz eta desbiderapen tipikoa antzeko mantentzen den bi kasuetan. Gazteen txioen luzeraren neurketekin jarraituz, mediana 8 tokenetan kokatzen da eta moda 6 tokenetan, ia 45.000 txio izanik 6 tokenetakoak. Mediana eta moda helduen taldearekin konparatuz gero, ikusten da mediak 6 tokenetako diferentzia daukala eta medianak 8 tokeneko diferentzia, alde nabarmena dagoela erakutsiz. Gazteen txioen luzerari erreparatuta, ikusi daiteke helduenak baino motzagoak direla oro har, pauta bat dagoela erakutsiz gazteen txioen luzeran.

HAP masterra



Irudia 10: Gazteen Txio pertsonalen tamainaren banaketa, txioen token kopuruaren 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 (Gaiak)	2.634.534	530.226
Euskarazko birtxioak (Harremanak)	2.421.058	400.448

Taula 16: Erauzitako datuak bizitza etaparen arabera.

5 Txiolarien ohiko gaiak identifikatu

Gazteen identifikatzailea garatuta dagoelarik, euskal txiolarien gaiak zeintzuk diren aztertzeari ekingo zaio. Aurreko atalean garatutako sailkatzaileari esker, gazte eta helduen gaiak argitara emateko aukera edukiko da, talde bakoitzaren gaiak modu independentean landuz.

Atal zehatz honetan, euskal txiolariak ze gairi buruz aritzen diren argitzea izango da asmoa. Hau da, euskal txiolarien elkarrizketa gairik ohikoenak detektatzea izango da asmoa. Datu honek lagunduko digu euskal komunitatearen agenda zein den ezagutzen, euskaldunen gairik errepikatuenak oinarri hartuta. Aldi berean, euskararen gaineko gogoeta egiteko ateak zabalduko dira, garai berrietara nola moldatzen ari den ezagutuz eta bereziki gazteek zertaz hitz egiten duten argituz. Honetarako, euskarazko 3 miloi txio baino gehiago erabiliko dira, helduak eta gazteak talde ezberdinduetan kokatuta.

Euskal txiolarien gaiak azaleratzeko atal hau lau zati ezberdinetan banatua egongo da. Lehenik eta behin, erauzketan lortu ditugun datuak nola aurreprozesatu diren azalduko da, hauen garbiketa nola eta zergatik egin den azalduz. Bigarren pausua, egitura gabeko datu mordo hau nola kudeatu azalduko da, euskal txiolarien gaiak ezagutuz *topic modeling* teknikari esker, *Latent Dirichlet Allocation* (LDA) algoritmoa nola aplikatuko den azalduz. Honekin batera, lortutako emaitzei interpretagarritasun erraz bat emateko asmoarekin, hauek interfaze grafiko baten laguntzaz argitara nola eman azalduko da ere. Azkenik, txiolari heldu eta gazteen gaiak azaleratuko dira eta bi taldeen arteko konparaketa egingo da.

5.1 Datuen aurreprozesaketa

Txioetatik informazio erabilgarria ateratzeko asmoarekin, lehenik eta behin datuen garbiketa burutuko da. Honekin, memoria aurreztearekin batera erroreak ekidingo dira, datuak garbitzeko emandako pausuak azalduz. Honez gain, ataza zehatz honetarako ere, lematizazioa erabili dugu gerora azalduko den moduan.

Datuen garbiketa: Lortutako datuak prozesatzeko orduan txioen testuetako hainbat karaktere ezabatzea erabaki da, testua sinplifikatzeko eta zarata ekiditeko. Lehenbizi, emotikono guztiak ezabatzeari ekin zaio, informazioa ekarri dezaketela uste den arren, python programazio lengoaiarekin hainbat errore sortzen ditu. Honekin batera, web-orrietarako esteka guztiak ezabatzea erabaki da ere, puntuazio arazoengatik hitzak tokenizatzerako orduan arazoak ematen zituen ere, batez ere esteka hauek zatitu egiten zirelako tokenizatzerako orduan. Horrez gain sinbolo eta puntuazio marka gehienak garbitzea erabaki da ere, prozesaketa arintzeko asmoarekin.

Hala ere, garbitu ez diren sinboloen artean '#' eta '@' daude, Twitter sare sozialean termino berriak sortzerako orduan sinbolo oso garrantzitsuak baitira, lehenengoa gaiak definitzerako orduan eta bigarrena pertsonak definitzerako orduan.

Lematizazioa: Datuen garbiketarekin batera ere, euskarazko hitzak lematizatzea erabaki

da, hau da, hitz bakoitzaren lema jaso eta bere atzizki zein aurrizki guztiak zokoratu. Euskara morfologikoki oso aberatsa den hizkuntza dugu, honek esan nahi du informazio asko atzizki zein aurrizkietan txertatuta doala. Beraz, badakigu lematizatzearen ekintza honekin informazio apur bat galduko dela, baina topikoen edo gaien detekziorako beharrezkoa dugu flexio guztiak erro berberarekin lotzea. Honela, hitzaren erroan doan informazioari emango diogu prioritatea, lortutako informazio mordoa homogeneizatuz.

Era honetan, euskarazko hitzak lematizatzeko, software berezi bat erabili behar izan da, normalean pakete ezagunak ingeleserako edo hizkuntza nagusituetarako bakarrik daudelako inplementatuta. Honela *"IXA pipes"* (Agerri et al., 2014) programaren barneko euskarazko lematizatzailea erabili da euskarazko hitzak lematizatzeko.

5.2 *Topic modeling* LDA erabiliz

Atal honen helburu nagusia euskal txiolarien gaiak identifikatzean datza LDA algoritmoaren bitartez. *Topic modeling*a, testu-meatzaritzaren alorrean maiz erabiltzen den tresna da. Tresna honekin, hitzak sailkatzeari ekingo zaio, antzeko hitzekin topiko orokorrago bat sortuz. Beste hitz batzuekin esanda, hitzak multzokatuko 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 LDA teknika erabiliko da, testuinguru bereko hitzekin topikoak sortuz.

Latent Dirichlet Allocation (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 modeloaren berezitasuna testuinguruan oinarritzen da, hitz baten taldekatzea bere inguruko hitzen arabera egingo du, inguruko hitzen ordena kontutan hartu gabe (BoW). Hortaz, modelo honi esker, kontzeptuak sailkatzeko esanahi semantikoa kontutan hartuko da. Hau da, antzeko esanahia duten hitzak multzo berdinetan taldekatuko dira eta, testuinguruari esker, sortutako multzoak definitzeko erraztasuna edukiko dugu.

Ekintza hau burutzeko lehenengo pausu moduan, dokumentuen egituratzea izango da. Txioen egitura bereziak, zaildu egiten du LDA ondo aplikatzea, hauek dokumentu laburregiak direlako eta antzekotasunak aurkitzea zaildu egiten duelako. Esan bezala, txioen izaera laburra kontutan hartuta, egituratze berezia egingo da, topic modelinga ondo burutu ahal izateko. Txioak erabiltzaileka multzokatuko dira, beste era batera esanda, dokumentu bakoitza erabiltzaile bakoitzaren txio pertsonalek osatuko dute. Honela, dokumentu bakoitzean erabiltzaile bakoitzaren txio guztiak jasota egongo dira. 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.

Adinaren araberako analisia egiteko, euskarazko 10 txio pertsonal baino gutxiago di-

tuzten erabiltzaileak kanpo geratu dira, sistemaren funtzionamendua zehatzagoa izateko, geroz eta testu kantitate handiagoa edukiz gero emaitza hobeak lortuko baitira. Honela, hasierako 7.980 erabiltzaileetatik, 7.087 erabiltzailek betetzen dute baldintza, 1.579 gazte eta 5.508 heldu.

LDA aplikatzerako orduan, aurrez zenbat gaiedo topiko identifikatu behar diren zehaztu beharko da, algoritmoak horrela funtzionatzen baitu. Era honetan, eredu bakoitzerako 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. Alde batetik, interpretagarritasunari dagokionez, errealitate sozialarekiko koherentzia edukitzea bilatu da ereduan, euskal gizartean aurkitu daitezkeen gaiekiko lotura izatea bilatuz. Bestalde, clusterren sakabanaketari dagokionez, ahalik eta eredu sakabanatuena lortzea izan da asmoa, taldeen arteko gainjartze ahalik eta txikiena bilatuz, taldeen arteko distantziak ezberdintasun kontzeptuala adierazten baitu modu berean. Ereduaren arabera (gazte/heldu), topiko kopuru desberdin bat aukeratu da, dokumentu kopurua handiagoa izan ahala topiko gehiago aukeratu daitezkeelako, informazioa kantitatea handiagoa baita. Hau da, *gazteen* ereduak topiko gutxiago dauzka informazio kantitatea txikiagoa delako. Honela, eredu bakoitzerako topiko kopuru ezberdinen arteko konparaketa burutu ostean, *helduen* eredurako 20 topiko erabiltzea erabaki da eta *gazteen* eredurako 12 topiko.

Erabili diren topiko kantitateak erabakitzeko prozesuan, ikusi ahal izan da, geroz eta topiko gehiago erabili geroz eta zehaztasun gehiago lortzen dela. Bestalde, topikoen kopurua txikitu ahala, orokorragoak diren terminoak lortu daitezke, sinplifikazioa lortuz. Hortaz, topiko kopuruaren aukeraketa egiteko prozesuan azaleratu da topiko kopuru egoki ugari daudela, topiko kopurua zehaztasuna edo orokortasuna lortzeko modu bat dela erakutsiz.

LDAren modeloa behin burututa, topikoak eta beren osagai diren terminoak jaso dira, hitz eta zenbaki mordoa lortuz. Bistaratze intuitiboago bat lortzeko asmoarekin, irudi grafikoak ematen dituen *LDAvis* metodoa erabiliko da, interpretazioan eta bistaratze lanetan oso lagungarria dena (Sievert eta Shirley, 2014). Helburu honetarako *pyLDAvis* paketea erabiliko da, *LDAvis* metodoa python programazio lengoaiara moldatuz. Pakete honi esker, topiko bakoitzeko cluster bat lortzen da, aldamenean termino bakoitzak clusterrean duen pisua ikusiz. Irudiaz gain, paketeak html fitxategi interaktibo bat ere eskuratzeko aukera eskeintzen digu, lamda hiperparametroa moldatzeko aukera ematen duelarik, topiko bakoitzaren terminoen agerpen pisua moldatzeko aukera emanez. Lambda hiperparametroa 1etik gertu dagoenean, topiko bakoitzean agerpen gehien dituzten terminoak edukiko ditugu. Bestalde, hiperparametro hau moldatuz, 0ra gerturatu ahala topiko konkretuetan soilik agertzen diren terminoak edukiko ditugu. Honela, lambda balio altu batekin topikoaren orokortasuna jaso ahalko da eta lambda balio txiki batekin topikoaren zehaztasuna.

LDA algoritmoaren bitartez hitzen clusterrak sortzeaz gain, bistaratze modu honek ere dimentsioen gutxitze bat burutzen dut. Lortutako topikoen clusterrak bi dimentsioetako espazio batean irudikatzen dira, horretarako Osagai Nagusien Analisia (*Principal Component Analysis*) aplikatzen duelarik. Dimentsioen gutxitze honek, interpretatzerako orduan

laguntza ematen digu, clusterrak dakarren informazioari bi ardatzetan daukan posizioa gehituz. Era honetan, cluster edo topiko antzekoak elkarrengandik oso gertu egongo dira eta oso ezberdinak direnak elkarrengandik urrun. Aldi berean, clusterren banaketa sakabanatua bada, topikoak egoki aukeratu direnaren seinale izango da, elkarrengandik ezberdintzen baitira. Horregatik guztiagatik, bistaratze interaktiboa ahalbidetzen duen algoritmo honi esker, LDA bitartez lortu ditugun topikoak interpretatzea errazagoa izango da. Hau da, topiko bakoitza gai batekin lotzeko erraztasuna emango digu gizarte zientzialarioi, topiko horiek eguneroko bizitzako gaiekin lotuz. Lambda hiperparametroa aldatzeko aukerari esker, gainera, topiko bakoitzaren orokortasunak zein zehaztasunak ikusi daitezke nahi dugun momentuan.

5.3 Emaitzen analisia

Esan bezala, behin topiko kopurua zehaztuta egonda eta emaitza *LDAvis* metodoari esker bistaratuz, topiko bakoitzaren identitatea argitzeari ekingo zaio. Topiko bakoitzaren identitatea, bere barneko hitz probableenek zehaztuko dute (Binkley et al., 2014) eta honi esker, euskal txiolarien gaiak zeintzuk diren argitu ahal izan da. Era honetan, bi ereduetarako, lortutako cluster bakoitzari izen bat jarri zaio zertaz hitz egiten den interpretatuz, cluster bakoitzean garrantzitsuak diren hitzak kontutan hartuz. Helburu horretarako, bistaratze interaktiboa erabili da, topiko bakoitza osatzen duten hitzak arakatuz eta topikoa definituz. Ataza honetarako bereziki lagungarria izan da lambda hiperparametroaren moldaketaren aukera, topiko bakoitzaren hitz garrantzitsuenak ikusteaz gain, aukera ematen baitu soilik topikoan agertzen diren hitzak bistaratzeko. Topikoak banan-banan definitu dira, 17. taulan ikusi daitekeen moduan, alde batean helduen gaiak eta bestean gazteen gaiak edukiz.

Helduen topikoak	**Hitzen %**
1 Elkarrizketa	% 10,5
2 Politika	% 10
3 Euskal txiolariak	% 6,9
4 Eskaintza kultural instit.	% 6,4
5 Administrazio publikoa	% 6,1
6 ETB	% 5,3
7 Txapelketak	% 5
8 Euskal presoak	% 4,9
9 Kultura	% 4,8
10 Herri mugimendua	% 4,8
11 Hezkuntza	% 4,3
12 Zientzia	% 4,1
13 Musika	% 3,9
14 Euskara	% 3,8
15 Kirola	% 3,8
16 Gipuzkoa	% 3,7
17 Euskarazko komunikabideak	% 3,5
18 Donostia	% 3,5
19 Nafarroa	% 2,7
20 Bizkaia	% 2,6

(a) Helduen gaiak.

Gazteen topikoak	**Hitzen %**
1 Gipuzkera	% 14,7
2 Sentimenduak	% 11,4
3 Bizkaiera	% 10,8
4 Kirola	% 9,9
5 Ekitaldi kulturalak	% 9,7
6 Zoriondu, eskertu	% 9,3
7 Bizitza kontatu	% 7,6
8 Bizkaiera formala	% 7,1
9 Gipuzkera formala	% 7,1
10 Euskal presoak	% 6,4
11 Athletic	% 3,3
12 Arrauna	% 2,7

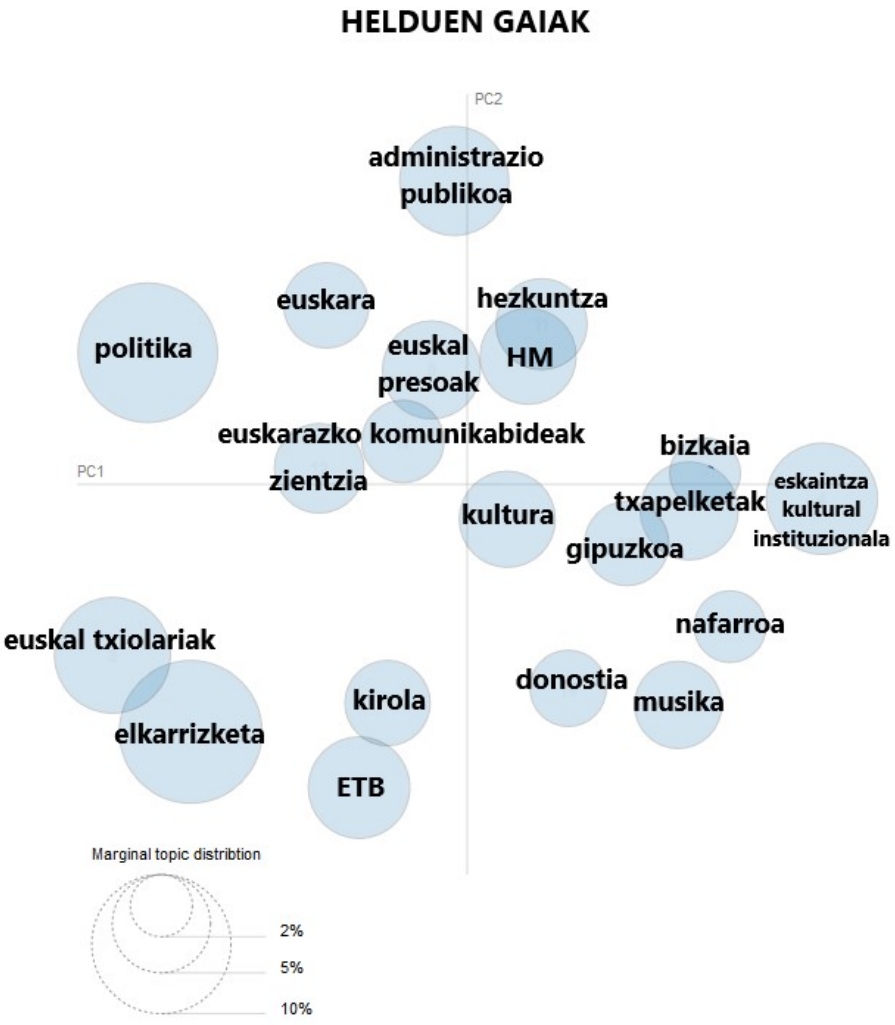
(b) Gazteen gaiak.

Taula 17: Twitterreko gaiak izendatuta.

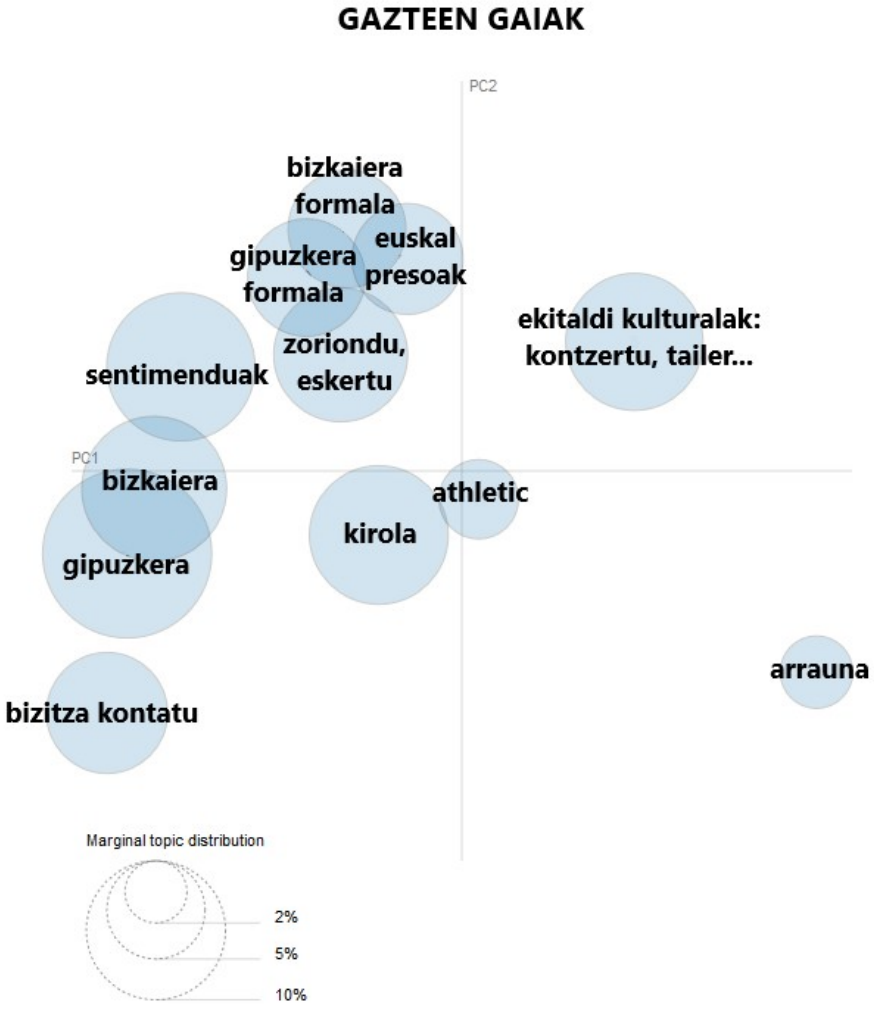
Helduen gaiei erreparatzen bazaie (17a taulan eta 11 irudian), gai asko politikarekin erlazionatuta daudela baiezta daiteke, adibidez, *politika*, *euskal presoak*, *herri mugimendua*, *hezkuntza* eta *euskara*. Honek erakusten du, Twitterreko euskal erabiltzaile helduek politikarekiko grina dutela. 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, **helduek batez ere gai politikoez edota gizarte gaiez aritzeko erabiltzen dute 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 (17b taulan eta 12 irudian), 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*, *Athtletic* 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 batez ere beren eguneroko bizitzan besteekin komunikatzeko**. Honekin ondoriozta daiteke, eguneroko bizitzan euskara erabiltzearen ondorio gisa, sare sozialetan ere euskara erabiltzeko joera naturala daukatela erabiltzaile hauek. Hau da, eguneroko bizitzan euskara erabiltzen bada, modu naturalean erabiliko dela sare sozialetan ere. Era honetan, ezagunak diren hainbat ideia modu kuantitatiboan islatu dira lan honetan.



Irudia 11: Helduen gaiak Twitterren.



Irudia 12: Gazteen gaiak Twitterren.

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

Ikerketari dagokion azken atal honetan, gazteen identifikatzailea garatuta dagoelarik, euskal txiolarien harremanak zeintzuk diren aztertzeari ekin zaio, birtxioetan oinarrituta. Garatutako formal/informal sailkatzaileari esker, iragarritako gazte eta helduen harreman sareak argitaratzeko aukera edukiko da, talde bakoitzaren harreman sarea modu independentean landuz.

Hortaz, laneko zati zehatz honetan, euskal txiolarien harreman sarea zein den erakustea izango da asmoa, horretarako erabiltzaile bakoitzaren birtxioak 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.

Harreman-sarea nola josten den ikustarazteko, euskarazko ia 3 milioi birtxio erabiliko dira, 400.448 gazteei dagozkienak eta 2.421.058 helduei dagozkienak. Elkarrekintzen sare bat lortuko dugu horrela, birtxio bakoitza elkarrekintza bat izanik.

6.1 Birtxioetan oinarritutako harreman sarea

Esan bezala, elkarrekintzetan oinarritutako harreman sare hau sortzeko, erabiltzaile bakoitzaren euskarazko birtxioak hartuko dira kontutan. Grafoa zuzendua izango da, hau da, birtxio hauen abiapuntua eta helburua kontutan hartuko dira, nork nor birtxiokatu duen kontutan hartzeko. Modu honetan, erlazioa nondik nora ematen den jakingo da eta hartzailearengan arreta berezia jarriko da.

Harremanen sare hau sortzeko, erauzketa fasean lortutako zerrendako erabiltzaileetatik aparte ere, askoz erabiltzaile gehiago lortu dira, birtxioak jaso dituzten erabiltzaileak ere sarera gehituz. Erauzketarako zerrendatik lortutako 5.508 erabiltzaile helduen 2.421.058 birtxioetatik habiatuta, 33.277 erabiltzaileren sarea lortzea lortu da. Lortutako 1.579 gazteetatik ere, 24.987 erabiltzaileren sarea eraikitzea lortu da, kasu honetarako 400.448 birtxio erabiliz. Honekin adierazi nahi da, sareak erraldoiak izango direla eta zerrendako erabiltzaileez gain, beste erabiltzaileak gehituko direla, beti ere zerrendako erabiltzaileen birtxioetan oinarrituz eta partekatutako edukia euskaraz izan baldin bada.

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 abiatuta sare bana eraiki da, heldu zein gazteentzat, *gephi* programa (Bastian et al., 2009) baliatuz eta arreta berezia bi puntutan jarriz:

HAP masterra

Azpitaldeak: Sare handi 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 saretik azpitalde edo komunitateak antzeman dira, erabiltzaileak birtxiokatu dituzten pertsonen arabera sailkatuz. Azpitalde hauetan sailkatzerakoan, komunitate desberdinak zeintzuk diren ikusi ahal izango da, erabiltzaile hauek nola eta zergatik harremantzen diren azaleratuz. Taldekatzeari esker ere, erabiltzaileen harremanak zertan oinarritzen diren interpretatzen lagunduko digu.

Nodoen tamaina: Grafoaren nodoen tamaina jasotako birtxioetan oinarrituko da, erabiltzaileak geroz eta birtxio gehiago jaso, orduan eta handiagoa izango da bere nodoa. Honek, euskal txiolarien komunitatearen baitako erabiltzaile erreferentzialak argitzen lagunduko digu, beti ere euskaraz argitaratutako edukia kontutan hartuz. Komunitate guztian erabiltzaile erreferentzialak zeintzuk diren argitzeaz gain, sortu diren azpitaldeak interpretatzeko baliagarria izango da. Honela, nodo garrantzitsuenei esker, azpitaldeen identitatea argituko da, erabiltzaile erreferentzialak baliatuz.

6.2 Emaitzen analisia

Behin eredu bakoitzarekin grafo bat sortu dela, eredu bakoitzari gainbegiratu bat emango zaio, azpitaldetan banaketa egin baino lehen, nodo garrantzitsuenak konparatuz bi ereduetan. Nodo garrantzitsu hauen konparaketa eginez, eredu bakoitzean garrantzitsuenak diren erabiltzaileak zeintzuk diren ikusi ahal izango da, hau da, arrakastatsuak diren erabiltzaileak zeintzuk diren ikusi ahal izango da. 18. taulan ikusi daitekeen moduan, gazteen zein helduen taldeetako nodorik garrantzitsuenak euskal komunikabideei dagokie (@berria, @argia, @naiz_info, @HamaikaTb, @topatu_eus...), Twitterren izaera komunikatiboa agerian utziz. Komunikabideekin batera, hauetako hainbat kazetari agertzen dira hauen artean, besteak beste @larbelaitz eta @axierL Argia astekariko kazetariak. Nodoen azaleko azterketa honetan, ezberdintasunei erreparatzen badiegu, badirudi helduentzat erreferentzialagoak direla albisteekin erlazionatutako komunikabideak (@eitbAlbisteak, @zuzeu) eta gazteengan, ostera, Ezker Abertzaleko pertsona (@ArnaldoOtegi, @jpermach) eta erakundeen (@ernaigazte) erreferentzialtasuna nabarmenagoa da. Erreferentziazko erabiltzaileen analisia burutu ostean, ikusi daiteke sare sozial honetan komunikabideen presentzia izugarrizkoa dela, informatzeko sare sozial bat dela ondorioztatu dezakegularik.

Helduen nodoak	Aldiz birtxiotua
@berria	3784
@argia	3478
@naiz_info	2513
@HamaikaTb	2334
@larbelaitz	2041
@boligorria	1746
@eitbAlbisteak	1741
@zuzeu	1710
@topatu_eus	1704
@axierL	1648

(a) Helduen nodo garrantzitsuenak.

Gazteen nodoak	Aldiz birtxiotua
@berria	925
@argia	788
@naiz_info	645
@larbelaitz	538
@topatu_eus	495
@ArnaldoOtegi	467
@ernaigazte	440
@HamaikaTb	425
@axierL	400
@jpermach	397

(b) Gazteen nodo garrantzitsuenak.

Taula 18: Erreferentziazko nodoak eredu bakoitzean.

Nodo garrantzitsuenei erreparatzeaz gain, eredu bakoitzaren berezitasuna diren nodoak argituko dira. Sailkapen honetarako, eredu bakoitzaren 100 nodo garrantzitsuenek hartzen duten posizioa hartu da kontutan. Eredu bakoitzean nodoak hartzen duen posizioa kontutan hartuta, rankingen arteko konparaketa bat egin da (gazteen nodo zerrenda erreferentzia moduan hartuz), eta kasu bakoitzeko rankingen arteko diferentziak lortu dira (Spearman, 1904). Era honetan, muturreko balio positiboekin gazteen nodo bereizgarriak lortu dira, honek esan nahi baitu nodo hauek garrantzi asko daukatela gazteen taldean eta gutxi helduen taldean. Bestalde, muturreko balio negatiboekin helduen nodo bereizgarrienak lortu dira, helduen nodo propioak agerian utziz. 19. taulan ikusi daiteke, gazteen artean, kirola (@iBROKI, @RealSociedad...) edo musikarekin (@ZuriHidalgo, @berritxarrak, @gaztea...) erlazionatutako nodoak daudela batez ere, gazteen harremanak aisialdiarekin erlazionatu daitezkeela erakutsiz.

Helduen nodo propioak
@Sustatu
@iPatxi
@EnekoitzEsnaola
@txargain
@xme64
@petxarroman
@zuzeu
@mikelgi
@samaravelte
@euskadi_irratia

(a) Gazteen nodoetatik gehien aldentzen diren helduen nodoak.

Gazteen nodo propioak
@iBROKI
@EsaldiakEuskara
@RealSociedad
@ZuriHidalgo
@berritxarrak
@IAbertzaleak
@euskarazEH
@gaztea
@MikelPagadi
@MeriLing1

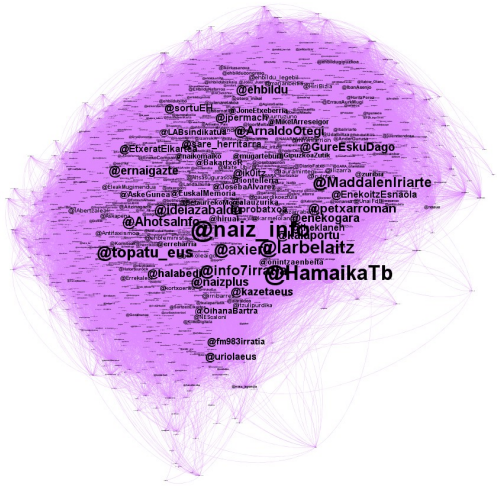
(b) Helduen nodoetatik gehien aldentzen diren gazteen nodoak.

Taula 19: Heldu eta gazteen arteko bereizketan oinarritutako nodoak.

Behin nodo garrantzitsuenen analisia burutu delarik, sakoneko azterketa batera igaroko da analisia. Kasu honetan, grafo bakoitza azpitaldeetan zatitu da, harremanak kontutan hartuz, taldeak nola osatzen diren ikusteko asmoarekin. Hau da, 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.

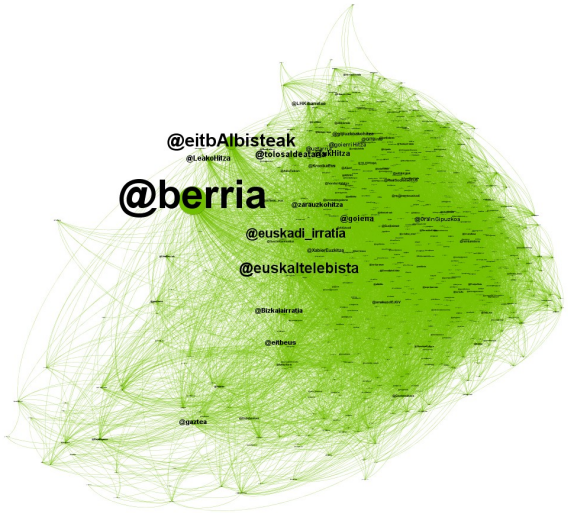
Helduen grafotik eratorritako azpitaldeei erreparatzen badiegu (20. taula eta 13-17 irudiak), ikusi daiteke gai nahiko ezberdinen inguruan harremantzen direla. Ezberdintasunak ezberdintasun, ikusi daiteke talde ezberdinek komunean daukaten ezaugarria, talde guztiek Euskal Herriko gaiekin erlazioa daukatela. 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, azpitalde ezberdinak gertaera hurbilekin erlazionatuta daudelarik. Honela, nahiko modu errazean erlazionatu ahal izan dira taldeak gai zehatzekin, jarraian ikusi ahal izango den moduan.

- *Ezker Abertzalea* (% 27,92): Azpitalde hau orientazio politiko konkretu bat daukaten pertsonen nodoez osatuta dago, zehazki Ezker Abertzalea osatzen duten erabiltzaileengatik osatua dago. 20. taulako lehenengo zutabean agertzen diren erabiltzaileak alde batera utzita, badaude nodo garrantzitsu ugari ezker abertzaleko militante edo erakundeenak direnak (@ArnaldoOtegi, @sortuEH, @jpermach, @LABsindikatua, @JosebaAlvarez...), nodo garrantzitsuenek baino hobeto azaltzen dutelarik talde honen idiosinkrasia. Beraz, talde nagusia, nodo guztien laurdena baino gehiago batzen duena, orientazio politiko jakin baten harremanei dagokio.



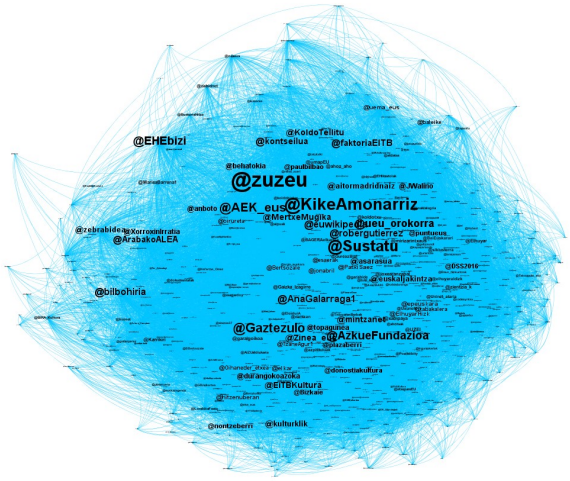
Irudia 13: Ezker Abertzalea (% 27,92).

- *Albisteak* (% 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.



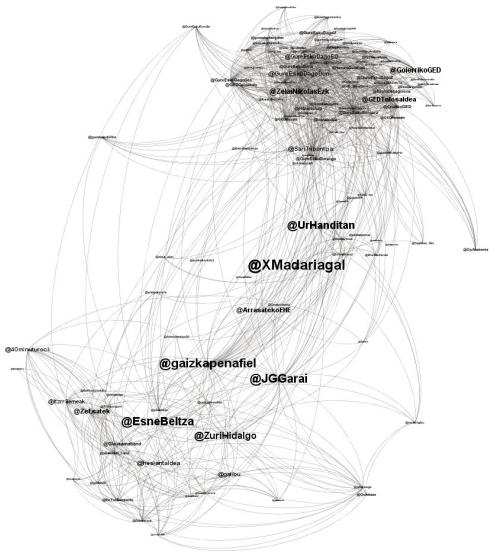
Irudia 14: Albisteak (% 23,77).

- *Euskara* (% 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.



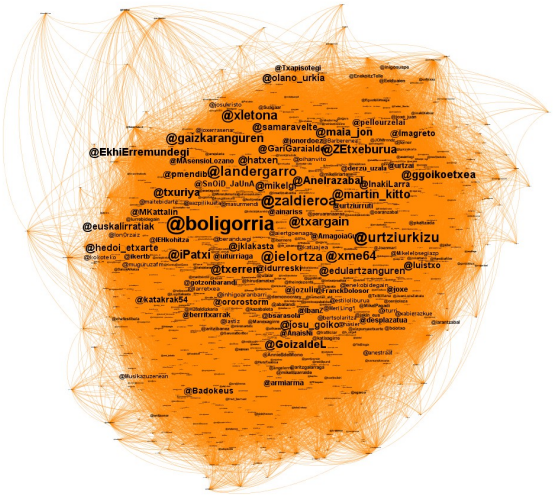
Irudia 15: Euskara (% 15,34).

- *Musika eta GED* (% 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 ostera Gure Esku Dago dinamikarekin zerikusia daukaten erabiltzaileekin erlazionatuta egongo litzateke (@GoierrikoGED, @GEDTolosaldea, @GureEskuDagoDon...).



Irudia 16: Musika eta GED (% 13,56).

- *Euskal Txiolariak* (% 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.



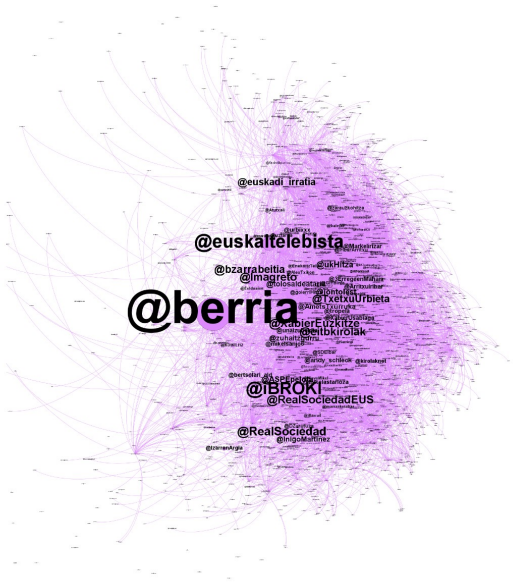
Irudia 17: Euskal Txiolariak (% 13,10).

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

Taula 20: Helduen azpitalde bakoitzeko nodo garrantzitsuenak.

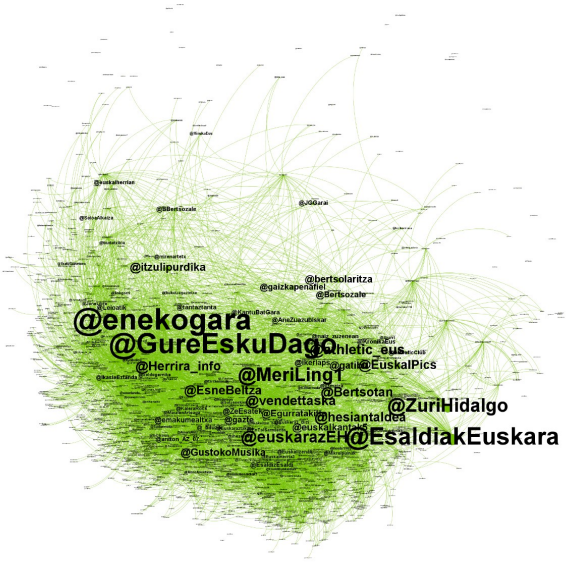
Gazteen grafotik eratorritako azpitaldeei erreparatuz gero (21. taula eta 18-22 irudiak), ikusi daiteke, antzekotasun eta ezberdintasunak daudela helduen grafoarekin konparatzerakoan. Helduen grafoarekiko antzekotasunei erreparatuz, ikusi daiteke, *Euskara*, *Ezker Abertzalea* zein *Albiste*en taldeak bi grafoetan azaldu direla. Bi grafoetan azaldutako azpitalde 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, azpitalde ezberdinak gertaera hurbilekin erlazionatuta daudelarik kasu honetan ere.

- *Kirolak* (% 21,61): Gazteen ereduko nodo gehienak barnebiltzen dituen azpitalde 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, @TxetxuUrbieta...) zein komunikabideak (@berria, @euskaltelebista, @eitbkirolak, @euskadi_irratia...). Beste behin ere, argi ikusi daiteke, kirolen taldea izan arren, komunikabideek badutela protagonismo erraldoia.



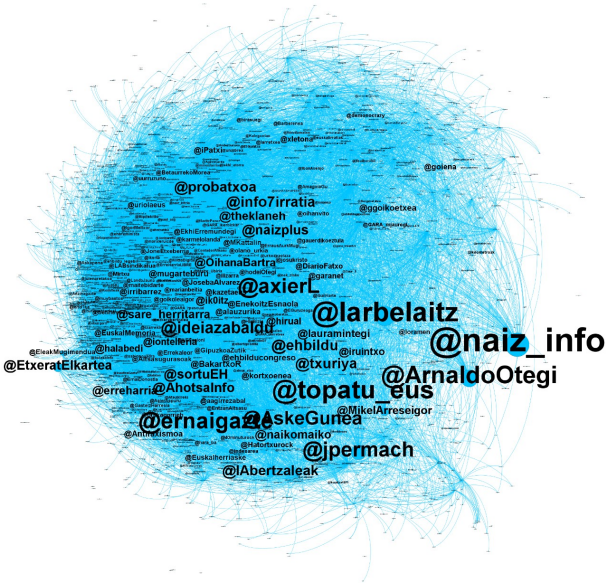
Irudia 18: Kirolak (% 21,61).

- *Euskara* (% 20,70): Bigarren azpitalde 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 azpitaldeetan ere gai honekin erlazionatutako komunitate bat aurkitu da, nahiz eta nodoak ezberdinak izan, gaia euskararekin erlazionatuta dagoela esan daiteke.



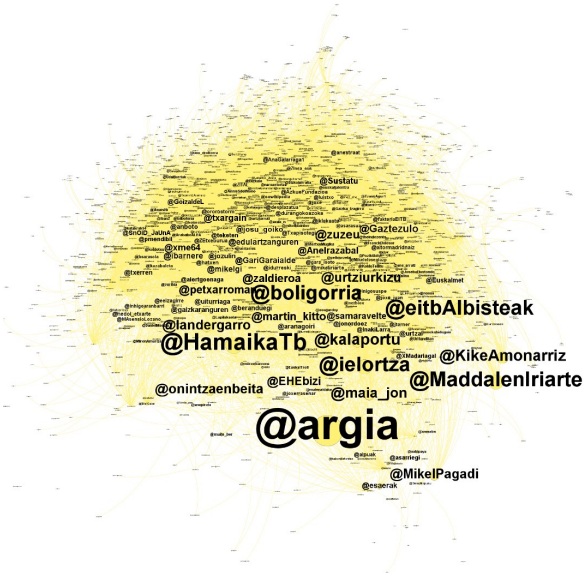
Irudia 19: Euskara (% 20,70).

- *Ezker Abertzalea* (% 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 norbanakoek (@ArnaldoOtegi, @lauramintegi...) azpitaldeari identitatea ematen diotelarik. Beste behin ere, ikusi daiteke helduen azpitaldeekiko antzekotasunak daudela, kasu zehatz honetan ere oso antzekoak direlarik.



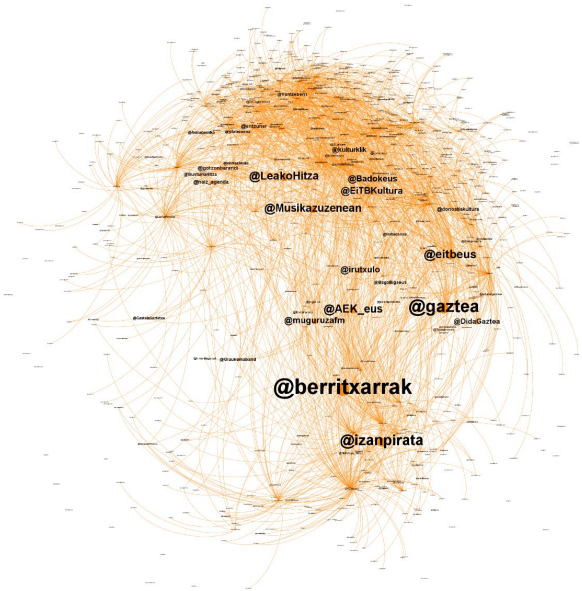
Irudia 20: Ezker Abertzalea (% 17,12).

- *Albisteak* (% 14,92): 4. talde honetan albisteekin erlazionatu ditugu bertako nodoak, nodo deigarrienak euskal komunikabideekin zerikusia baitaukate (@argia, @HamaikaTb, @eitbAlbisteak, @zuzeu, @Gaztezulo). Kasu honetan ere, azpitalde hau eta helduen ereduan azaldutako azpitaldeak erlazioa daukate, bi eredu ezberdinen artean paralelotasunak ematen direla berriz ere erakutsiz.



Irudia 21: Albisteak (% 14,92).

- *Musika* (% 11,35): Azkeneko azpitaldea, nahiko talde anitza izan arren, musikarekin erlazionatutako aktore ezberdinekin erlazionatu ditzakegu nodo garrantzitsueneta-ko batzuk. Aktore horien artean musika hizpide duten komunikabideak (@gaztea, @DidaGaztea), musika taldeak (@berritxarrak, @muguruzafm, @Glaukomaband) eta baita disketxeren bat ere bai (@BagaBigaeus) edukiko genituzke.



Irudia 22: Musika (% 11,35).

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

Taula 21: Gazteen azpitalde bakoitzeko nodo garrantzitsuenak.

Behin bi ereduetako azpi taldeak interpretatu eta definitu direnean, bien arteko konparaketa egin da, horretarako 22. taula lagungarri izanik. 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 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 azpitaldeetan ikusi ahal izan dugu, harremantzeko modua komunitate konkretuen baitan ematen dela, komunitate horien gaia nahiko erraz intuitu daitekeelarik. Harreman azpitaldeen 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.

Helduen azpitaldeak	**Nodo kopurua**
Ezker Abertzalea	% 27,92
Albisteak	% 23,77
Euskara	% 15,34
Musika eta GED	% 13,56
Euskal Txiolariak	% 13,10

(a) Helduen harremanen komunitateak.

Gazteen azpitaldeak	**Nodo kopurua**
Kirolak	% 21,61
Euskara	% 20,70
Ezker Abertzalea	% 17,12
Albisteak	% 14,92
Musika	% 11,35

(b) Gazteen harremanen komunitateak.

Taula 22: Harremanen komunitateak eredu bakoitzean.

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.

Datuen erauzketaren inguruko balorazioa burutzerakoan, esan daiteke datu-iturri berri bat ireki dela Twitterreko informazioa erauzteko garatutako teknikari esker. Erauzketa teknika honi esker, datu kantitate handiak (Big Data) lortu dira oso kostu txikiarekin. Horrez gain, gazteen ingurunera hurbiltzea lortu da, hauen errealitatea hobeto ezagutzeko lehenengo pausua emanez. Hala ere, datu-iturri berri honek ere baditu bere mugak, Twitterreko erabiltzaileengan mugatzen dela ikerketa. Hobekuntza moduan, datu-iturri berriak gehitu daitezke beste sare sozial batzuen erauzketa burutuz, adibidez Instagram, Snaptchat edota Facebook erauziz, edukia publiko jarriko balute. Baina, teknika honekin sare sozialen erabiltzaileen informazioa soilik lortuko litzateke eta, gaur egun behintzat, populazioaren zati handi bat kanpo geratuko litzateke. Horrez gain ere, Twitterrek berak jarritako mugak (erabiltzaileko 3200 txio gehienez jota eta 15 erabiltzaile bakarrik 15 minuturo) ere kontutan hartu behar dira, lana dezente mantsotu egiten baitute muga horiek. Datu-iturri berri honen mugak alde batera utzita, sistema arrakastatsu bat garatu dela onartu beharra dago, ia 8.000 euskal erabiltzaile erauztea lortu baitira, 10 milioi txio baino gehiago geureganatuz.

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, 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 txikiekin ere sailkapena ondo egiten duela esan beharra dago. Etorkizunerako, egokia izango litzateke sailkapena adinaren arabera burutzea, horretarako corpus berri bat etiketatu beharko litzatekeela argi edukiz. Hala ere, *IXA pipes* dokumentu sailkatzaileari esker, etiketatutako corpus txikiekin ere sailkapen egokiak burutu daitezke, sistema hornitzen duten egiturarik gabeko datu multzoetan oinarritutako hitzen irudikapenei esker. Etiketatutako corpus txikiekin ere ondo sailkatzen duen sistema bati esker, ikertzaile txikiek aukera gehiago daukate atazak modu arrakastatsuan garatzeko, corpus handiak etiketatzea lan eta kostu handia baitira.

Behin gazte eta helduen taldeak ezberdinduta daudelarik, hauen gai ohikoenak zeintzuk diren azaleratu eta hauen harremantzeko modua zein den argitu da hein batean. 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 argitzen saiatu gara. 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 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.

8 Bibliografia

Agerri, R., Bermudez, J., & Rigau, G. (2014, May). *IXA pipeline: Efficient and Ready to Use Multilingual NLP tools.* In LREC (Vol. 2014, pp. 3823-3828).

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

Agerri, R., & Rigau, G. (2018). *Language Independent Sequence Labelling for Opinion Target Extraction.* Accepted in Artificial Intelligence Journal.

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

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

Bauman, Z. (2015). Modernidad líquida. Fondo de cultura económica.

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.* Journal of statistical mechanics: theory and experiment, 2008(10), P10008.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). *Class-based n-gram models of natural language.* Computational linguistics, 18(4), 467-479.

Castells, M., Gimeno, C. M., & Alborés, J. (2005). *La sociedad red* (Vol. 1). Alianza.

Cesare, N., Grant, C., & Nsoesie, E. O. (2017). *Detection of user demographics on social media: A review of methods and recommendations for best practices.* arXiv preprint arXiv:1702.01807.

Chen, S. F., & Goodman, J. (1999). *An empirical study of smoothing techniques for language modeling.* Computer Speech & Language, 13(4), 359-394.

Clark, A. (2003, April). *Combining distributional and morphological information for part*

of speech induction. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1 (pp. 59-66). Association for Computational Linguistics.

Collins, M. (2002, July). *Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms.* In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 1-8). Association for Computational Linguistics.

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

Gamallo, P., Pichel, J. R., & Alegria, I. (2017). *From language identification to language distance.* Physica A: Statistical Mechanics and its Applications, 484, 152-162.

González Bermúdez, M. (2015). *An analysis of twitter corpora and the differences between formal and colloquial tweets.* In Proceedings of the Tweet Translation Workshop 2015 (pp. 1-7). CEUR-WS. org.

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).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* In Advances in neural information processing systems (pp. 3111-3119).

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.* PloS one, 12(8), e0183537.

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2014). *How Old Do You Think I Am? A Study of Language and Age in Twitter.* In ICWSM.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). *Computational sociolinguistics: A survey.* 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. 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 In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Roesslein, J. (2009). tweepy Documentation. *Online] http://tweepy. readthedocs. io/en/v3*, 5.

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).

Spearman, C. (1904). *The proof and measurement of association between two things.* The American journal of psychology, 15(1), 72-101.

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.