Sentimenduen analisirako lexikoen sorkuntza

Iñaki San Vicente eta **Xabier Saralegi**

Elhuyar Fundazioa
Osinalde Industrialdea 3,
20170 Usurbil
{i.sanvicente,x.saralegi}@elhuyar.com

Laburpena

Testuetan adierazten diren sentimendu eta iritziak automatikoki aztertzeko oinarrizko baliabideak dira polaritate-lexikoak. Euskaraz, horrelako teknologia garatzeko ahaleginak oso urriak izan dira orain arte. Artikulu honetan lexiko horiek modu automatikoan sortzen hiru bide aztertu dira: beste hizkuntzetan dauden lexikoak itzultzea, testu-corpusetatik erauztea, eta WordNet moduko ezagutza base eleaniztunen gainean sentimenduak markatzea. Emaitzek erakusten dute metodo hauek baliagarri direla polaritate-lexiko eraginkorrak hutsetik modu azkar batean eta adituen ahalegin handirik gabe sortzeko.

Hitz gakoak: Sentimenduen Analisia, Polaritate-lexikoak

Abstract

Polarity lexicons are a basic resource for analyzing the sentiments and opinions expressed in texts in an automated way. Very little work has been done on this regard for Basque. This paper explores three methods to automatically construct polarity lexicons: translating existing lexicons from other languages, extracting polarity lexicons from corpora, and annotating sentiments in WordNet likeMultilingual Lexical Knowledge Bases. Results show that these methods are useful for creating lexicons from scratch fast and with little effort from human experts.

Keywords: Sentiment Analysis, Polarity Lexicons

1 Sarrera eta motibazioa

Iritzi-erauzketa eta sentimenduen analisiaren motibazioa domeinu komertzial eta politikoak aztertzeko aplikazioen beharretik dator. Aplikazio horien helburua gizartearen sentimendu eta jarrerak era automatikoan jarraitzea litzateke, berri, foro, eta abarren bidez. Zein da gizarteak Ukrainiako gatazkari buruz duen iritzia? Zein da jendeak marka batekiko duen harrera? Eta modelo zehatz bat kaleratu ondoren? Testuetatik abiatuz iritziak eta emozioak identifikatuko lituzkeen sistema bat oso baliagarria litzateke horrelako galderei erantzun ahal izateko.

Sentimenduen analisiaren alorrak azken urteetan izugarrizko bultzada izan du, hainbat jardueratan oso interesgarriak baitira, hala nola zaintza teknologikoan, marketin alorrean produktu zein enpresen inguruko iritzia ezagutzeko, pertsonen gaineko izen ona aztertzeko, gai gatazkatsuen inguruko erreakzioak antzemateko, eta abar.

Ikerketa-ildo hori azkenaldian horrenbeste hazi izana Web 2.0ren etorrerarekin lotu behar da. Internet berriak erabiltzaileei edukiak sortzeko ahalmena eman die. Orain arte, produktu, erakunde edo gai baten inguruan gizartearen iritzia inkesten eta arreta zerbitzuen bidez bildu izan da, baina horrek erabiltzailea eta enpresaren zuzeneko harremana eskatzen zuen. Erabiltzaileok, baina, ez ditugu bide horiek askotan erabiltzen, askoz ohikoagoa da gure iritzia lagunartean adieraztea. Orain gutxi arte informazio hori eskuratzea oso zaila zen enpresentzako, baina gaur egungo Internetek horrelako informazioa gordetzen du eta eskuragarri jartzen du edozeinentzako. Iritzi-erauzketak datu masa erraldoi horretatik informazioa

erauzi eta prozesatzeko aukera ematen du, komunitateak gai zehatz baten inguruan une batean duen pentsamoldea inferitu dezakegularik.

Testu-unitate baten polaritatea identifikatzeko baliabide nagusiena polaritate-lexikoak dira. Polaritate-lexikoak hitz bakun edo hitz anitzeko unitateen zerrendak dira, non sarrera bakoitzaren a prioriko polaritatea adierazita dagoen. Horrela, Mandela hil da aste honetan"esaldia, polaritate-lexiko batek "hil"negatiboa dela jasota badu, esaldi hori negatiboa dela ondorioztatu genezake.

Tamalez, horrelako lexikoak sortzea ez da lan makala. Eskuz sortzeak kostu handia du, eta lexikoaren estaldura mugatua da. Lan honek polaritate-lexiko horiek sortzeko metodo automatikoak zein erdi automatikoak aztertzen ditu. Gure helburua euskarazko lexikoak sortzea da; horrek, baina, gure aukerak mugatzen ditu, eskura ditugun baliabideak urriak baitira. Ez dugu polaritate informazioa markatuta duen dokumentu multzorik, eta, ondorioz, ezin dugu ikasketa automatikoko estrategiarik erabili.

Artikulua honela dago antolatuta: hurrengo atalean polaritate sailkapenaren inguruko literatura laburbilduko da, arreta berezia eskainiz polaritate-lexikoen sorkuntzari. Ondoren, euskarazko lexikoak sortzeko aztertu ditugun teknikak deskribatuko dira. 4 atalak sortutako lexikoen egokitasuna aztertzeko burutu dugun ebaluazioa eta bere emaitzak aurkeztuko ditu, eta azkenik, 5 atalean lortutako ondorioak eta etorkizuneko lan ildoak zehaztuko dira.

2 Artearen Egoera

2.1 Bi hitz Polaritate Sailkapenaren inguruan

Polaritate sailkapenaren inguruko literatura laburpen oso egokiak argitaratu dira (Liu *et al.*, 2012; Pang eta Lee, 2008). Polaritatearen sailkapenari aurre egiteko literaturan proposatutako metodoak bi multzo handitan banatzen dira; Alde batetik, gainbegiratutako metodoak edo adibideetan oinarritutakoak aurki ditzakegu eta, bestetik, ez-gainbegiratuak edo ezagutza linguistikoan oinarrituak.

Metodo ez-gainbegiratuak erregela linguistikoetan eta polaritate-lexikoetan oinarritzen dira. Polaritate-lexikoak eskaintzen duen hitz mailako polaritatea baliatuz, esaldi edo dokumentu mailako polaritatea kalkulatzen da.

Hurbilpen gainbegiratua polaritatea markatua duen erreferentzia edo entrenamendu corpus batetik polaritate ezberdinak bereizten dituzten ezaugarrien araberako sailkatzaileak ikastean datza. Ikasitako sailkatzaile automatikoa jasotako testu berriei polaritatea esleitzeko gai da. Lagin hori eskuz etiketatzeak suposatzen duen kostu altua ekiditeko asmoz, testuak sailkatuta dituzten iturrietara jotzen da. Horrela, produktu, filma edo bestelako elementuen kritika sailkatuak dituzten webguneetara jotzen da entrenamenduko corpusak biltzera. Ohikoa da Amazon[1], tripAdvisor[2] edo IMDB[3] moduko webguneetatik lortzea testuak, kritika bakoitzak polaritate sailkapen bat baitu.

Metodo gainbegiratuetan oinarritutako estrategia arrakastatsuagoa (Pang *et al.*, 2002),(Chaovalit eta Zhou, 2005) eta literaturan erabiliena bada ere, sistema komertzial askoren nukleoa polaritate-lexikoek osatzen dute. Izan ere, sailkatzaile automatikoak entrenatzeko behar diren datuak domeinu eta hizkuntza ezberdinetan lortzea oso zaila da. Bestalde, lexikoak estrategia gainbegiratuarekin konbinatuz gero emaitzak hobetu daitezkeela frogatu da (Wilson *et al.*, 2005).

2.2 Polaritate-lexikoen sorkuntza

Literaturan polaritate-lexikoak sortzeko proposamenak hiru multzo nagusitan antolatu daitezke: eskuz sortutako lexikoak (Stone *et al.*, 1966; Taboada *et al.*, 2011), Ezagutza Base Lexikaletan (EBL) oinarritutako metodoak (Kamps *et al.*, 2004; Liu eta Singh, 2004; Kim eta Hovy, 2004; San Vicente *et al.*, 2014) eta corpusetan oinarritutako metodoak (Hatzivassiloglou eta McKeown, 1997; Turney eta Littman, 2003; Mihalcea *et al.*, 2007) .

Hizkuntza handietarako eskuz sortutako polaritate-lexiko oso ezagunak daude, hala nola, General Inquirer(Stone *et al.*, 1966), OpinionFinder(Wilson *et al.*, 2005), edo SO-CAL(Taboada *et al.*, 2011). Horiek

[1]http://www.amazon.com
[2]http://www.tripadvisor.com
[3]http://www.imdb.com

sortzeko behar diren giza baliabideak oso altuak direla eta, horietako batzuk erdi automatikoki sortuak dira, eta ondoren eskuz zuzenduak. Horrela, dagoeneko hizkuntza baten existitzen diren polaritate-lexiko eta entrenatze corpusak beste hizkuntza batzuetarako berrerabiltzea aztertzen duten lanak ere badaude. Mihalcea *et al.* eta Perez-Rosas *et al.* ikertzaileek baliabideak ingelesetik errumanierara 2007 eta espainierara 2012 itzulpen automatiko bidez itzultzea bezalako estrategiak aztertzen dituzte, hurrenez hurren. Lexikoen kasuan atal txiki batek bakarrik mantentzen du polaritatea itzuli ostean. Nabarmena egiten da itzulpen anbiguoak tratatzeko beharra.

Corpusetan oinarritutako metodoek, nolabaiteko polaritate anotazioa behar dute lexikoak eraucteko. Bi hurbilpen nagusi daude multzo honetan: lehena, polaritate ezaguna duten hitz batzuetatik abiatuta, corpusetan hitz horien semantikoki antzekoak diren hitzak aurkitzean datza (Turney eta Littman, 2003). Bigarrena, polaritatea markatuta duen corpus batean oinarrituz, positiboenak zein negatiboenak diren hitzen zerrendak lortzea (Saralegi eta San Vicente, 2012)

Azkenik, EBLetan oinarritutako metodoen funtsa da polaritate ezaguna duten hazi-hitz multzo txiki batetik abiatuta, EBLak eskaintzen dituen hitzen arteko loturak baliatuz, hasierako polaritate horiek hitz berrietara hedatzea. (Hatzivassiloglou eta McKeown, 1997) lanean hazi-lexiko bat hedatzen dute and/but motako konektoreen bidez. Mohammad *et al.* lanean tesauro bat baliatzen dute hasierako hazi-hitz positibo eta negatibo batzuen sinonimoak markatu eta polaritatea zabaltzeko. WordNetek (WN) (Fellbaum, 1998) eskaintzen dituen erlazio semantikoez osatutako grafoetan zehar kontzeptu batzuen polaritatea hedatzea ere oso estrategia erabilia da (Esuli eta Sebastiani, 2006; San Vicente *et al.*, 2014).

3 Lexikoak sortzeko metodoak

Aurreko ataletan aipatu dugun bezala, polaritate-lexikoak testuen polaritatea detektatzeko baliabiderik garrantzitsuenetako bat dira. Lexiko horiek sortzerakoan, euskaraz aplikatu daitezkeen kostu baxuko hiru estrategia aztertu ditugu lan honetan: (i) beste hizkuntza batean existitzen diren lexikoak euskarara itzultzea edo proiektatzea; (ii) corpusetatik polaritatea adierazten duten terminoak automatikoki erauztea; eta (iii) EBLetan dauden hitzen polaritatea markatzea.

3.1 Proiekzioa

Polaritate-lexikoak sortzeko aukera zuzena dirudi beste hizkuntza batean dagoeneko sortuta dagoen lexiko bat itzultzea hiztegi elebidunak baliatuz. Estrategia honek, hala ere, bere arazoak ditu; itzulpen prozesuaren ondorioz ematen den kalitate-galerari aurre egin behar dio. Baina, bestalde, baliabide independientea da printzipioz, eta edozein datu-sortaren gainean erabil daiteke, hau da, domeinuarekiko menpekotasunik gabekoa da. Corpusetan oinarritutako lexikoak, aldiz, garatutako corpusaren menpekoak dira, eta, ondorioz, beste domeinu batzuetan aplikatzean errendimendu-galera bat izan lezakete.

Proiekzioari dagokionez, gaztelaniazko $ElhPolar_{es}$ (Saralegi eta San Vicente, 2013) polaritate-lexikoa itzuli dugu. Itzulpena burutzeko, Elhuyar Fundazioaren gaztelania-euskara[4] hiztegia erabili dugu, 173.931 itzulpen bikote dituena. Gaztelaniazko sarrera bakoitzeko lehen 5 itzulpenak hartu dira Lex_{pr} sortuz. Ondoren, bi adituk euskarazko hitzen polaritatea berrikusi dute, polaritate zuzena anotatuz eta polaritaterik gabeko itzulpenak baztertuz. Ezadostasunak eztabaida bidez konpondu dira Lex_{przuz} hiztegia lortuz. 1 taulak prozesuan sartutako lexikoen estatistikak ematen ditu.

	#sarrera	**#positibo**	**#negatibo**
$ElhPolar_{es}$	5.195	1.892	3.303
Lex_{pr}	11.413	4.934	6.479
Lex_{przuz}	9.299	3.911	5.388

1 Taula: ElhPolar jatorrizko eta itzulitako lexikoen sarrera estatitstikak.

Zuzenketa kostua Orokorrean, eskuzko lanari leporatzen zaion arazo nagusia bere kostu altua da. Lan honetan lexikoa eskuz zuzentzeak suposatu duen ahalegina neurtu dugu. Horretarako, bi adierazle hartu ditugu kontutan:

[4] http://hiztegiak.elhuyar.eus

- Zuzenketa abiadura: minutuko zenbat hitz zuzentzen diren.
- Emankortasuna: Polaritate hitzak lortzeko abiadura, hots, minutuko gure lexikoan zenbat polaritate-hitz gehitzen diren.

1 Irudiak bi zuzentzaileen arteko batezbesteko zuzenketa abiadura eta emankortasun denborak erakusten ditu, hautagai tarteak adierazita (Corpusetan oinarritutako lexikoarekin alderatzeko, lehen 5.000 hautagaien datuak bakarrik daude irudian islatuta).

Osotara, 36 ordu behar izan dira euskarara automatikoki itzulitako lexikoa Lex_{pr} zuzentzeko, hau da, Lex_{przuz} sortzeko. Horrek esan nahi du minutuko bataz beste 5,3 hitz zuzendu direla. Orokorrean esan daiteke, lan handia exijitzen duela, hiztegi bidezko itzulpenak ez-ohiko ordain asko sortzen baititu, eta horien esanahia hiztegi eta corpusetan bilatu behar izateak zuzentzaileen lana nabarmen moteldu du.

3.2 Corpusetan oinarritutako lexikoak

Hurbilpen horren oinarrian dagoen helburua da antzematea zeintzuk diren polaritate jakin bateko (positibo ala negatibo) dokumentuetan agertzeko joera duten hitzak. Hitz horiek topatzeko elkartze-neurriak erabili ohi dira (Kilgarriff, 2001; Rayson eta Garside, 2000).

Estrategia egokiena litzateke anotatutako corpus bat hartuta adibide positiboa eta negatiboak banatzea. Tamalez, ez dago euskaraz horrelako anotaziorik duen corpusik, eta erdibideko hurbilpen batera jo behar izan dugu, testu subjektiboak eta objektiboak bereizita dituen corpus bat baliatuz. Horrelako corpus bat eraikitzeko estrategia merke bat hartu dugu Berriako artiku bilduma batetik -C_{Berria}- abiatuta: Iritzi-artikuluak subjetibotzat hartu dira, eta gainerakoak objektibotzat (C_{Berria}) (Saralegi *et al.*, 2013). 2 Taulak corpus horren neurriak eta erauzketaren datuak azaltzen ditu. Elkartze-neurri gisa, dokumentu subjektiboekiko hitz lotuenak identifikatzeko, egiantz-arrazoia (LLR) (Dunning, 1993) erabili dugu.

Aurretik azaldutako metodologia jarraituz hitz subjektiboak erauzi ditugu C_{Berria} corpusetik, polaritate-hitz hautagaiak alegia. Ondoren, eskuz esleitu zaie polaritatea zerrenda horretatik subjektibotasun neurri altuena duten 5000 hitzei. Eskuzko anotazio hori bi pertsonen artean burutu da, ezadostasunak eztabaida bidez konpondu direlarik.

Corpus	**#hitz**	**#dok**	**#hitz subj.**	**#dok subj.**	**#hitz obj.**	**#dok obj.**
C_{Berria}	20.402.121	75.892	3.810.857	13.325	16.591.264	62.567

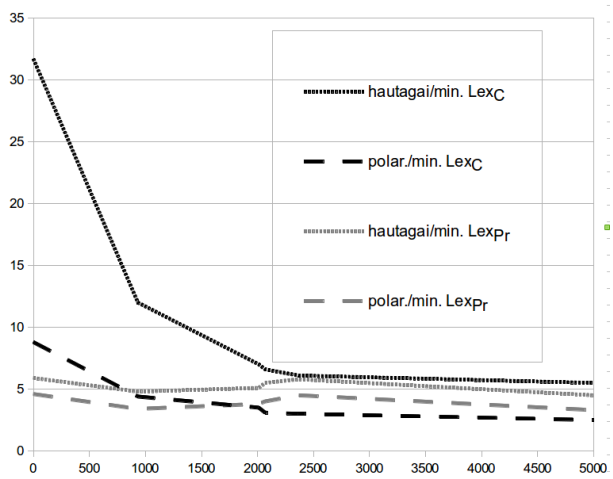
2 Taula: Berriako corpusaren eta bertatik ateratako datu-sorten estatistikak.

Zuzenketa kostua Proiekzioaren kasuan ez bezala, hemen polaritateak hutsetik esleitu behar ditu anotatzaileak, baina, bestalde, hitzak corpus batetik ateratakoak diren heinean erabiliagoak dira, eta, beraz, ezagunagoak eta errazagoak polaritatea anotatzeko. Kostua neurtzeko 3.1 ataleko adierazle berdinak erabili ditugu. 1 irudian bi anotatzaileen denboren batezbestekoak ageri dira. Irudiak erakusten duen moduan, abiadura askoz altuagoan esleitzen da polaritatea LLRren araberako zerrendaren hasierako hautagaien artean, zerrendan aurrera joan ahala zalantzazko kasuak areagotzen baitira. Era berean, zenbat eta aurrerago joan orduan eta polaritate hitz gutxiago topatzen dira LLRk adierazitako subjektibotasun-maila baxuagoa delako.

Osotara, 10 ordu behar izan dira C_{Berria} corpusetik erauzitako 5000 hautagaiko zerrenda lantzeko, polaritate esleipena eta zalantzazko kasuen eztabaida barne. Lortutako lexikoak (Lex_c) guztira 1.659 sarrera ditu (959 negatibo eta 691 positibo).

3.3 EBLetan oinarritutako lexikoak

EBLek eskaintzen dituzten lotura semantikoak baliatzea polaritate-lexikoak sortzeko oso estrategia erabilia da literaturan. Gure kasuan, San Vicente *et al.* autoreek 2014an proposatutako metodoa erabili dugu euskarazko lexikoak sortzeko. Metodo honek hiru elementu behar ditu polaritate-lexiko bat sortzeko: (i) EBL baten grafo errepresentazioa; (ii) informazioa grafoan zehar hedatzeko algoritmo bat; eta(iii) algoritmoa abiarazteko hazi multzo bat, hitz edo kontzeptuak (synsetak).



1 Irudia: Eskuzko hiztegien zuzenketaren errendimendu datuak.

QWN-ppv-k WNen gisako EBL batean dauden erlazio semantikoak grafo baten bidez errepresentatzen ditu. Grafo horretako erpinak EBL kontzeptuak dira eta ertzak kontzeptuen arteko erlazio semantikoak (e.g., sinonimia). Grafo horren gainean polaritate informazioa zabaltzeko Personalized PageRank algoritmoaren Agirre eta Soroaren UKB hurbilpena 2009 aplikatzen da. PageRank algoritmo ezagunak grafo baten erpinen ranking bat sortzen du, erpinek grafoaren egituran duten garrantzian oinarrituz. Jatorrizko algoritmoan erpin guztiak pisu berdinarekin hasieratzen dira, eta aldiz, UKBk erpin batzuek hasieran garrantzia handiagoa izan dezaten ahalbidetzen du. Gure kasuan, polaritate jakina (positiboa edo negatiboa) duten hazi-hitz edo hazi-kontzeptu batzuei pisu handiagoa emanez hasieratzen dugu algoritmoa, eta hala, polaritatearen arabera ordenatutako ranking bat sortuko du algoritmoak.

Ondoren, lan honetan erabilitako konfigurazioa azaltzen da. Ematen diren parametroen balioak (San Vicente *et al.*, 2014) lanean ateratako ondorioetatik eratorriak dira.

EBL gisa MCR (Agirre *et al.*, 2012) erabili dugu. MCR-ek hainbat hizkuntzetako WNak bateratzen ditu kontzeptu mailan, tartean euskarazkoa. Honek abantaila ematen digu, izan ere, Ingelesezko WNak askoz erlazio semantiko gehiago eskaintzen ditu beste hizkuntzek baino, polaritate informazioa zabaltzeko aukera gehiago ematen digularik. MCRren grafo errepresentazioa sortzean erlazio MCReko sinonimia eta antonimia erlazioak dituzten grafo bana sortu dugu. 2 grafo eta 2 hazi multzo izanda (positibo eta negatiboak) 4 hedapen lortzen ditugu. Amaierako lexikoa lortzeko, hurrengo formula erabiltzen da:

$$Lex_{qwn-ppv} = Sinonimia_{pos} + Antonimia_{neg} - Sinonimia_{neg} + Antonimia_{pos} \quad (1)$$

EBLaren beste errepresentazio batzuk erabiltzea badago, esaterako erlazio guztiak dituen grafo bakarra. Sinonimia eta antonimia bidezko grafoek, erlazio gutxiago izanik, hedapen murritzagoa egiten dute alde batetik, baina, bestetik, doitasun handiagoko hedapenak sortzen dira.

Haziei dagokienez, hemen ere aukera dugu MCR barruan dauden edozein hizkuntzetako haziak zein kontzeptuak hazi gisa erabiltzeko. Gure kasuan kontzeptuak erabili ditugu. Hasierako hazi-kontzeptu horien multzoa osatzeko, ingelesezko 14 hitz positibo eta negatiboko zerrenda batetik abiatu gara (Turney eta Littman, 2003), eta horien inguruko kontzeptuak bildu ditugu, WNeko erlazio hauen bidez lotutakoak: antonymy, similarity, derived-from, pertains-to and also-see. Hemen ere hasierako hazietatik hainbat pausutara dauden kontzeptuak bildu ditzakegu. Gure kasuan gehienez ere 8 pausuko distantziara zeuden kontzeptuak bildu ditugu.

Metodo honekin lortutako lexikoak ($Lex_{qwn-ppv}$) 1.132 sarrera ditu, 565 positibo eta 567 negatibo.

4 Ebaluazioa

Hiztegien ebaluazioa burutzeko gainditu beharreko oztopoak ez dira gutxi. dokumentu edo esaldi multzo etiketaturik ez izateak sailkapenean oinarritutako ebaluazio estrintseko bat burutzeko ezintasuna dakar, eta ez dago eskuz sortutako euskarazko polaritate-lexikorik, beraz ebaluazio intrintsekorik ere ezin burutu eskuz lagin bat aztertzen ez bada.

Azkenean, ebaluazio estrintseko baten alde egin dugu. Horretarako testerako esaldi multzoa guk markatu dugu. 4.1 atalak deskribatzen du datu-multzo horren sorrera.

Polaritatea sailkatzeko batez besteko polaritatea hitzen kontaketa bidez kalkulatzen da. Oso sistema sinplea da, baina, gure helburua lexikoen egokitasuna aztertzea izanik, beste aspektu batzuk kontutan hartzen dituzten sistema aurreratuagoak erabiltzea baztertu da. Horrela, erregeletan oinarritutako sailkatzailea inplementatu dugu. Sailkatzaileak t testu baten polaritatea kalkulatzeko L_{eu} polaritate-lexiko baten informazioa hartzen du oinarritzat. Testu osoaren subjektibotasun maila testuko hitzen polaritateen batezbestekoa da, 2 ekuazioak adierazten duen moduan.

$$Pol(t) = \sum_{w \in t} bal(w)/\#w \tag{2}$$

non $bal(w) = L_{eu}(w)$ w hitzak polaritate-lexikoan duen polaritatea den (-1 ala 1) eta #w t testuaren hitz kopurua den. $Pol(t) > 0$ balio batek testua positiboa dela adierazten du, $Pol(t) =< 0$ balio batek aldiz negatiboa.

4.1 Test datuak

Testerako esaldiak bi domeinutatik hartu ziren: domeinu periodistikotik (Gara eta Berria egunkarietatik) eta musika kritiken domeinutik (Gaztezulo aldizkaritik). Guztira 193 esaldira bildu eta polaritatearen arabera anotatu ziren: positiboak eta negatiboak. Polaritate neutroa zutenak baztertu ziren.

Domeinua	**Positibo**	**Negatibo**	Guztira
Musika kritikak	%87,27	%12,73	55
Kazetaritza	%25,36	%74,64	138
Guztira	%43	%57	193

3 Taula: Ebaluaziorako datu-sorten estatistikak.

4.2 Emaitzak

4 Taulak erakusten ditu lexiko ezberdinek lortutako emaitzak datu-sorta ezberdinetan. Doitasun (Acc.) eta kategoria bakoitzaren araberako Fscore (Fpos/Fneg) neurriak erabili dira lexiko bakoitzaren errendimendua neurtzeko. Emaitza aztertzen baditugu, corpusetan oinarritutako lexikoak lortzen ditu emaitza onenak neurri eta datu-sorta guztiei dagokienez. Espero bezala eskuz zuzendutako lexikoek guztiz automatikoek lortutako emaitzak gainditzen dituzte.

Lexikoa	**#sarrera**	**Kazetaritza**			**Musika kritikak**			**Osotara**		
		Acc.	Fpos	Fneg	Acc.	Fpos	Fneg	Acc.	Fpos	Fneg
Lex_{pr} (Itzulpena)	11.413	0.63	0.41	0.73	0.62	0.73	0.32	0.63	0.57	0.67
Lex_{przuz} (Itzulpena)	9.299	0.6	0.47	0.68	0.73	0.83	0.35	0.64	0.63	0.64
Lex_{c} (Corpus)	1.659	**0.75**	0.5	0.84	**0.78**	0.86	0.5	**0.76**	0.7	0.8
$Lex_{qwn-ppv}$ (EBL)	1.132	0.67	0.21	0.79	0.22	0.25	0.19	0.54	0.23	0.68

4 Taula: Datu-sorta ezberdinen gainean sortutako polaritate lexikoek lortutako asmatze tasak, polaritate kategoria bakoitzeko zehaztuta.

Musika arloko emaitzetan, nabarmentzekoa da esaldi negatiboen sailkapenean orokorrean izan den errendimendu baxua. Hau azaldu daiteke, batetik, esaldi negatiboen kopuru oso txikia delako (7), eta,

bestetik, esaldi horiek aztertuta, ikusi delako hitzen polaritateaz kanpoko beste fenomeno linguistiko batzuk erabakigarriak direla polaritatea zehazterakoan. Halaber, aipatzekoa da $Lex_{qwn-ppv}$ lexiko automatikoaren emaitza kaxkarra. Bere errendimendua oso baxua izan da bereziki esaldi positiboen kasuan.

Kazetaritza arloko emaitzei erreparatzen, espero gabeko emaitza bat izan dugu, izan ere, itzulitako lexikoen artean, Lex_{przuz} zuzendutako lexikoak emaitza okerragoak ditu Lex_{pr} zuzendu gabekoak baino. Emaitzek erakusten dute esaldi negatiboen detekzioan dagoela gakoa (Ikus 4 taulan Fneg zutabea).

5 Ondorioak eta Etorkizuneko norabidea

Lan honetan euskarazko polaritate-lexikoak sortzeko hiru bide aztertu ditugu: beste hizkuntzetako lexikoak proiektatzea; corpusetatik erauztea eta EBLetan oinarrituta hutsetik sortzea. Domeinu orokorreko lexikoak sortu dira, eta, hala, beren egokitasuna bi domeinutako datuen gainean ebaluatu da.

Lortutako emaitzen arabera, metodorik egokiena hiztegiaren errendimenduari dagokionez, corpusetan oinarritutako lexikoen erauzketa erdi-automatikoa litzateke. Gainera, metodo horrek eskatzen duen eskuzko ahalegina ez da oso handia (10 ordu). Lexikoa oso handia ez bada ere, corpusetan oinarritua egoteak hitz subjektibo erabilienak barne hartzen dituela bermatzen du metodoak.

EBLetan oinarritutako metodo erabat automatikoak emaitza global okerrenak lortzen ditu, musikakritiken alorrean duen errendimendu baxuagatik. Hala ere, albisteen alorrean itzulitako lexikoek baino emaitza hobeak lortzen ditu. Kontutan izanda kostu baxueneko metodoa dela aintzat hartzeko alternatiba da. Halaber, ikusi dugu euskarazko WNak ez duela adjektiboen informaziorik apenas. polaritatearen detekzioan adjektiboak oso garrantzitsuak dira, eta beraz, gure ustea da lortutako lexikoen kalitatea nabarmen hobetuko litzatekeela informazio hori edukita. Interesgarria litzateke eusWN euskarazko adjektiboekin aberasteko bideak aztertzea.

Azkenik, itzulitako lexikoa da aztertutako metodoen artean lan gehien eskatzen duena (36 ordu), eta gainera emaitzek ez dut eskuzko lan horren onura garbirik erakusten. Emaitzak sakonago aztertu beharra dago arazoa non dagoen argitzeko, metodo hau erabat baztertu aurretik. Izan ere, ezin da ahaztu baliabide aurretik metodo honek hiztegi elebidun bat besterik ez duela eskatzen, besteek EBLak eta corpus etiketatuak eskatzen dituzten bitartean.

Erreferentziak

Agirre, Aitor González, Egoitz Laparra, German Rigau, eta Basque Country Donostia. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference*, p. 118.

Agirre, Eneko, eta Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Chaovalit, P., eta L. Zhou. 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, p. 112c.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19.61–74.

Esuli, A., eta F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, 417–422, Genoa, Italy.

Fellbaum, Christiane. 1998. *WordNet*. Wiley Online Library.

Hatzivassiloglou, V., eta K. R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 174–181.

Kamps, Jaap, Maarten Marx, Robert J. Mokken, eta Maarten De Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of LREC 2004*, Lisbon, Portugal.

Kilgarriff, A. 2001. Comparing corpora. *International journal of corpus linguistics* 6.97133.

Kim, Soo-Min, eta Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of Coling 2004*, 1367–1373, Geneva, Switzerland. COLING.

Liu, H., eta P. Singh. 2004. ConceptNet: a practical commonsense reasoning toolkit. *BT Technology Journal* 22.211226.

Liu, Xiaohua, Furu Wei, Ming Zhou, eta Microsoft QuickView Team. 2012. QuickView: NLP-based tweet search. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, p. 1318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mihalcea, R., C. Banea, eta J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Annual Meeting of the Association for Computational Linguistics*, volume 45, p. 976.

Mohammad, S., C. Dunne, eta B. Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 599–608.

Pang, B., eta L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2.1–135.

Pang, Bo, Lillian Lee, eta Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 79–86. Association for Computational Linguistics.

Perez-Rosas, Veronica, Carmen Banea, eta Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, eta Stelios Piperidis, Istanbul, Turkey.

Pérez-Rosas, Verónica, Carmen Banea, eta Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *LREC*, 3077–3081.

Rayson, Paul, eta Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora - Volume 9*, WCC '00, p. 16, Hong Kong, China. Association for Computational Linguistics. ACM ID: 1117730.

San Vicente, Iñaki, Rodrigo Agerri, eta German Rigau. 2014. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, 88–97.

Saralegi, Xabier, Iñaki San Vicente, eta Irati Ugarteburu. 2013. Cross-lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In *Computational Linguistics and Intelligent Text Processing*, ed. by Alexander Gelbukh, volume 7817 of *Lecture Notes in Computer Science*, 96–108.

Saralegi, Xabier, eta Iñaki San Vicente. 2012. Tass: Detecting sentiments in spanish tweets. In *Proceedings of the TASS Workshop at SEPLN*.

Saralegi, Xabier, eta Iñaki San Vicente. 2013. Elhuyar at TASS2013. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2013)*, 143–150, Madrid.

Stone, P., D. Dunphy, M. Smith, eta D. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge (MA): MIT Press.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, eta Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37.267–307.

Turney, P., eta M. Littman. 2003. Measuring praise and criticism: Inference of semantic oreintation from association. *ACM Transaction on Information Systems* 21.315–346.

Wilson, Theresa, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, eta Siddharth Patwardhan. 2005. OpinionFinder. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* -, 34–35, Vancouver, Canada.

Wilson, Theresa, Janyce Wiebe, eta Paul Hoffman. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 347354.