Subjektibitatearen azterketa euskarazko testuetan

Egilea: Iñaki San Vicente Roncal
Tutorea: Kepa Mirena Sarasola Gabiria

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2012ko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia, Elektronika eta Telekomunikazioak.

Laburpena

Lan honetan, euskarazko testuetan subjektibitatea detektatzeko teknikak lantzen dira. Subjektibitatea detektatzeko lexikoak sortzeko garaian, beste hizkuntza batean sortutakoak itzultzea edo lexikoak corpusetatik automatikoki sortzea hobea den aztertu da. Horrez gain, subjektibitatea aztertzeko balio dezaketen zenbait ezaugarri aztertzen dira, hitzen kategoria gramatikala bereziki. Dokumentuen nahiz esaldien subjektibitatea landu da, test datu-sorta ezberdinen gainean.

Abstract

This work addresses the problem of detecting subjectivity in Basque tests. We adopt an unsupervised approach based on sujectivity lexicons, and compare translated subjectivity lexicons with lexicons created from scratch using automatic techniques. In addition, other features that can help detecting subjectivity are also analised, such as POS tags. Subjectivity detection is evaluated both at document and at sentence level, with regard to different test sets.

Eskerrak

Eskerrak eman nahi dizkiot Elhuyar I+G unitatean kide dudan Xabier Saralegiri, alor honetan egindako ikerketetan bidelagun dudalako, eta lan hau burutzeko eman dizkidan aholkuengatik. Eskerrak Xabier Rojori ere, ikerketa honetan erabilitako programazio-kodearen zati bat berak idatzia delako.

Gaien aurkibidea

1 Proiektuaren definizioa

Erakundeentzat guztiz baliagarria da jakitea haien jardueren aurrean gizarteak nola erantzuten duen. Enpresen lehiakortasuna hobetzeko ere, oso garrantzitsua da besteek produktuen zein enpresaren gainean zer iritzi duten jasotzea. Erakunde politikoei interesatzen zaie jakitea gizarteak nola ikusten dituen haiek aurrera eramaten dituzten politikak. Hori orain arte inkesten eta arreta-zerbitzuen bidez egin izan da, baina horrek erabiltzailearen eta enpresaren zuzeneko harremana behar zuen.

Aitzitik, erabiltzaileok ez ditugu bide horiek askotan erabiltzen, askoz ohikoagoa baita gure iritzia lagunartean adieraztea. Orain gutxira arte informazio hori eskuratzea oso zaila zen enpresentzako, eta ahalegin handia eskatzen zuen.

Gaur egun, aldiz, Internetek horrelako informazioa gordetzen du, eta edozeinentzako eskuragarri jarri. Iritzi-erauzketak informazio-masa erraldoiak erauzi eta prozesatzeko aukera ematen du, komunitateak gai zehatz baten inguruan une batean duen pentsamoldea inferitu dezakegularik.

Iritzi-erauzketaren eta sentimendu-analisiaren alorrek azken urteetan izugarrizko bultzada izan dute, hainbat jardueratan oso interesgarriak baitira, besteak beste: zaintza teknologikoan; marketinean, produktuen zein enpresen inguruko iritzia ezagutzeko; pertsonek izen ona duten aztertzeko; eta gai gatazkatsuen inguruko erreakzioak antzemateko.

Ikerketa-ildo horiek azkenaldian horrenbeste hazi izana Web 2.0 sortzearen ondorioetako bat da. Internet berriak erabiltzaileei eduki-kontsumitzaile huts izatetik eduki-sortzaile izatera pasatzeko ahalmena eman die, eta gaur egun, edozeinek du aukera bere pentsamoldea argitaratzeko. Horien adibide dira blogak, Facebook moduko sare sozialak, Twitterren gisako microblogging zerbitzuak eta produktuei buruzko iritziak emateko guneak (e.g., RottenTomatoes[1] filmeei buruzko iritziak emateko, edo Tripadvisor[2] jatetxe eta hotelei buruzko iritziak partekatzeko). Tresna horien erabilera masiboa bilakatzen ari da.

Euskaldunok ere parte hartzen dugu zerbitzu horietan. Euskarazko tuitak biltzen dituen eu.umap.eu zerbitzuak 800.000[3] tuitetik gora atzeman zituen 2010eko azken bi hilabetetan bakarrik. Ikus daitekeenez, iritziak detektatu eta analizatzeko teknologia oso baliagarria litzateke euskal erakundeentzat.

Bestalde, Informazioaren Gizartean eleaniztasunak duen garrantzia handituz doa egunez egun. Informazioa zenbait hizkuntzatan sortzeko, jasotzeko eta trukatzeko eskaera gero eta handiagoa da, eta hizkuntz teknologiek zeresan handia dute eskakizun horri erantzuteko erabiliko diren tresnen garapenean.

Iritzia erauzteko garaian ere, ia ezinbestekoa da hainbat hizkuntzatan dagoen informazioa tratatzeko gai izatea, informazio hori ez baitago hizkuntza bakarrean normalean. Ezin

[1] http://www.rottentomatoes.com/
[2] http://www.tripadvisor.es
[3] http://www.codesyntax.com/eu/cs-tailerra/files/umap-2010eko-estatistikak

dugu ahaztu euskal gizartea elebiduna dela, eta, beraz, edukiak euskaraz nahiz gaztelaniaz sortzen ditugula (eta gero eta gehiago baita ingelesez ere). Horrenbestez, etorkizunera begira, oso garrantzitsua litzateke sortzen den teknologia eleaniztasuna tratatzeko gai izatea, baita iritzien detekzioaren kasuan ere.

1.1 Subjektibotasuna eta Polaritatea

Wilson et al. (2005b)-ek adierazpen subjektibo gisa definitzen dituzte iritzia, emozioa, jarrera, espekulazioa, etab. adierazten duten hitzak edo esamoldeak. Horiek izendatzeko termino orokorra "egoera pertsonal" litzateke. Quirk et al. (1985)-ek egoera pertsonala honela definitzen dute:

> *"Behaketa edo egiaztapen objektiborik egin ezin denean. Adibidez, egiazta daiteke pertsona batek jainkoa existitzen dela esan duela, baina ez pertsona horrek jainkoa existitzen dela sinesten duela."*

Zentzu horretan, sinesmena pertsonala da. Beraz, subjektibotasunaren detekzioa iritziak, emozioak eta bestelako subjektibotasun-formak adierazteko testu-unitateak (hitzak, esamoldeak, esaldiak, paragrafoak,...) egitatezko informazioa objektiboki adierazteko erabiltzen direnetatik bereiztean datza.

Hala ere, lan hori ez da batere samurra, subjektibotasuna ez baita balio boolear huts batetara mugatzen. Wilson (2008) egoera pertsonala noiz adierazten den eta zer ezaugarri dituen aztertzeko lanak subjektibotasun-analisian sartzen ditu. Ezaugarri horiek barne hartzen dute egoera pertsonala nork adierazten duen, intentsitatea, zer iritzi edo jarrera adierazten den, nor edo zeri buruzko iritzia den, eta abar.

Lan honetan, dokumentu edo esaldi batek iritzirik adierazten duen detektatzera mugatuko gara. Dokumentuen kasuan, sailkapen hori egitea errazagoa dirudi: dokumentu osoaren asmoa iritzia adieraztea den erabaki behar da. Azter liteke iritzi hori testu osoan zehar banatuta dagoen edo parte batean bakarrik ematen den (amaieran, adibidez), baina hori lan honen mugetatik kanpo geratzen da.

Esaldiren bat iritzia adierazten duen erabakitzeko orduan, zalantza gehiago sor daitezke. Quirk-en adibidearekin jarraituz, demagun ondorengo esaldia sailkatu behar dugula: *"Ionek esan du sinesten duela jainkoa existitzen dela"*. Esaldi honetan egoera pertsonal bat dago: *"Ionek jainkoa existitzen dela sinesten du"*. Era berean, gertakari objektibo bat ere badago: *"Ionek X esan du"*.

Zeri eman garrantzi gehiago? Lan honetan horrelako esaldiek iritzizkotzat jo ditugu. Horren arrazoia da esaldi horiek iritzi bat adierazten dutela, nahiz eta iritzi hori hirugarren batek adierazi. Hala ere, iritziaren intentsitatea landu beharko bagenu, *"Ionek esan du sinesten duela jainkoa existitzen dela"* eta *"Ionek jainkoa existitzen dela sinesten du"* esaldiek intentsitate-maila ezberdina dutela esango genuke.

Iritzia duten testu-unitateak identifikatzea baliagarria bada ere, are baliagarriagoa litzateke iritzi horiek (oso) positiboak edo (oso) negatiboak diren detektatzea. Horrela, gai baten inguruan ematen diren iritziak sailka genitzake, estatistikak atera eta aldeko nahiz kontrako iritzien inguruko laburpenak osatu. Hain zuzen ere, polaritatea testu-unitate batek sentimendu positiboa edo negatiboa adierazten duen identifikatzean datza.

Iritzi-erauzketak eta sentimenduen analisiak subjektibotasunaren detekzioa eta polaritatearen estimazioa biltzen dituzte. Ataza biak dira garrantzitsuak iritzi-erauzketako aplikazioetan; esaterako, iritziak bildu eta laburtzeko orduan eta produktuak konparatzean. Iritzi-erauzketan lanean diharduten ikertzaileek erakutsi duten bezala, bi urratsez osatutako hurbilpena onuragarria izan daiteke kasu askotan: lehenik, instantzia subjektiboak eta objektiboak bereizten dira, eta, ondoren, instantzia subjektiboak polaritatearen arabera sailkatu (Yu eta Hatzivassiloglou, 2003; Pang eta Lee, 2004; Wiebe et al., 2005; Kim eta Hovy, 2004).

Hortaz, euskarazko testuetan subjektibotasuna hautematea lehenengo pauso egokia da iritzia erauzteko sistema bat eraikitzeko. Gainera, testua prozesatzeko aplikazio ugaritan erabili dute dagoeneko subjektibotasunaren azterketarako metodoren bat, adibidez: testutik ahotserako sintetizatzaileetan (Alm et al., 2005), testuaren analisi semantikorako aplikazioetan (Esuli eta Sebastiani, 2005; Wiebe eta Mihalcea, 2006), sareko foruetako eta berrietako sentimenduen denbora-lerroen jarraipenean (Bollen et al., 2010; Balog et al., 2006), eta galderei erantzuteko sistemetan (Yu eta Hatzivassiloglou, 2003).

1.2 Proiektuaren helburua

Lan honen helburua iritziaren erauzketarako sistema eleaniztun baten lehen pausuak ematea da. Gai honen gainean gero eta ikerketa gehiago argitaratzen ari dira, baina ingelesaren inguruan gehienak, eta eleaniztasunaren eremuan oraindik lan gutxi egin da. Euskararekin lotutako ikerketarik ez dugu ezagutzen.

Iritzien detekzioaren alorrean, euskarak ohiko arazo bati aurre egin behar dio: ikerketalan gehienak ingelesaren inguruan egin dira, eta, ondorioz, ingeleserako hainbat baliabide daude sortuta (polaritatea markatuta duten lexikoiak, adibidez). Tamalez, gainerako hizkuntzetan baliabide horiek ez dira horren ugariak, eta, are gehiago, hizkuntza gehienek (euskara kasu) ez dute horrelako baliabiderik.

Proiektu honetan dokumentuetako subjektibotasunaren detekzioan zentratuko gara. Gure helburua izango da testu batek iritzia azaltzen duen identifikatzea. Horretarako, literaturak ikasketa automatikoan oinarritutako teknikak (Riloff eta Wiebe, 2003) eta erregeletan oinarritutakoak proposatu ditu (Yu eta Hatzivassiloglou, 2003). Literaturan lehen hurbilpenak emaitza hobeak eman baditu ere (Riloff eta Wiebe, 2003; Chaovalit eta Zhou, 2005), ikasketa automatikoko sailkatzaileak entrenatzeko datu nahikoa biltzea ez da lan erraza.

Hori dela eta, guk lan honetan bigarren hurbilpena aztertuko dugu. Hurbilpen horren baliabide nagusia subjektibotasun-lexikoa da. Baliabide hori eskuratzeko bideak aztertuko ditugu: dagoeneko existitzen diren baliabideak euskarara egokitzea ala baliabide horiek hutsetik sortzea.

Batetik, baliabide horiek euskaratzeak esan nahi du itzulpen-prozesu bat burutzea, eta horrek baliabideen kalitatea txiki dezake neurri batean, itzulpenaren kalitatearen arabera. Bide horretatik, ingeleserako sortuta dagoen lexiko bat hiztegi bidez automatikoki itzuliko dugu.

Bestetik, baliabidea hutsetik sortzeak haren kalitatea bermatzen du, baina kostua handiagoa da. Guk subjektibotasun-lexikoa sortuko dugu euskarazko corpus objektibo bat eta corpus subjektibo bat gurutzatuz. Horrek kostua murriztuko du, prozesu automatikoak baliabidearen kalitatea guztiz bermatzen ez duen arren.

Gure sistema garatu eta ebaluatzeko, subjektibotasun-informazioa atxikita duen adibide-sorta bat behar da. Guk dakigula, ez dago horrelako baliabiderik euskararako; beraz, guk sortu ditugu. Dokumentuen subjektibotasuna sailkatuta duten euskarazko bi corpus sortu dira, kazetaritzaren alorrekoak. Bat tresna garatzeko eta subjektibotasun-lexikoak erauzteko prozesuan erabili da, eta bestea, berriz, ebaluazio-prozesuan, sistemaren sendotasuna aztertzeko.

Bi sorta horiek baliagarriak izan daitezke etorkizunean, ikasketa automatikoan oinarritutako sistemekin esperimentatzeko. Horiez gain, esaldien arabera sailkatuta dagoen adibide-sorta bat ere prestatu da, gure metodoa testu-unitate laburragoak tratatzeko gai den ebaluatzeko.

Garatutako sistema aipatutako bildumen gainean ebaluatuko da, zehaztasun-, doitasun- eta estaldura-datuak kalkulatuz.

Txosten hau honela dago antolatuta hemendik aurrera: hurrengo atalean iritzi-erauzketaren alorrean egin diren lanen laburpen bat eskaintzen da, subjektibotasunaren detekzioari eta haren inguruko baliabideei arreta berezia eskainiz. atalean lan honetan erabili diren baliabideak aurkezten dira, erabilitako corpusak, baliabide lexikalak eta tresna linguistikoak. Ondoren subjektibitatearen detekzioa burutzeko erabili dugun estrategia deskribatzen da, baliabideak sortzeko erabili ditugun metodoak eta gure sistema ebaluatzeko diseinatu ditugun esperimentuak azalduz. 5 atalean egindako esperimentuen emaitzak aurkezten dira eta horien irakurketa egiten da. Amaitzeko egindako ikerketak utzitako ondorioak aipatzen dira, eta etorkizuneko ikerketa bideratzeko proposamenak egiten dira.

2 Aurrekariak

1.1 atalean aipatu dugun bezala, hainbat ikertzaileren arabera testu-unitateen polaritatea sailkatu aurretik subjektiboak diren ala ez erabaki behar da. Pang eta Lee (2004) autoreek

doitasuna % 82,8tik % 86,4ra hobetzen dute polaritatearen sailkapenean, subjektibotasunaren detekzioari esker.

Subjektibotasuna detektatzeko lehen ikerketak 90eko hamarkadaren amaieran egin ziren, testuetan subjektibotasuna adierazten duten hitzak erabiliz (Hatzivassiloglou eta McKeown, 1997; Bruce eta Wiebe, 1999). Zerrenda horiek eskuz (Pang et al., 2002) ala automatikoki (Hatzivassiloglou eta McKeown, 1997; Yu eta Hatzivassiloglou, 2003; Maks eta Vossen, 2012) osa daitezke, hizkuntza subjektiboa egokien ordezkatzen duten elementuak topatzeko neurri estatistikoak erabiliz.

Subjektibotasunaren detekzioa burutzeko, bi hurbilketa nagusi proposatu dira literaturan: ikasketa automatikoan oinarritutako metodo gainbegiratuak (Yu eta Hatzivassiloglou, 2003) eta erregeletan oinarritutakoak (Riloff eta Wiebe, 2003).

Erregeletan oinarritutako hurbilketak subjektibotasun-lexikoak erabiltzen ditu ezagutzaren oinarri gisa, eta, ondoren, metodo estatistikoak erabiltzen ditu testuen subjektibotasun-maila kalkulatzeko. Riloff eta Wiebe (2003) ikertzaileek proposatutako metodoa subjektibotasuna adierazten duten hitz-zerrendetan oinarritzen da.

Zerrenda horietako hitzek bi mailatan banatuta dute subjektibotasuna: altua eta baxua. Sistemak subjektibotzat jotzen du esaldi bat, subjektibotasun-maila altua duten bi hitz baino gehiago topatzen baditu. Das eta Bandyopadhyay (2009b) autoreek bengalerarako erregeletan oinarritutako sistema bat aurkezten dute, hainbat ezaugarri konbinatzen dituena: maiztasuna, lexikoetan agertzea (SentiWordNet), POS informazioa, hitzen ordena, etab.

Ikasketa automatikoak, berriz, sailkatzailea entrenatzeko datuak behar ditu, hots, subjektibotasuna etiketatua duen corpus bat. Pang eta Lee (2004) egileek Naive Bayes eta SVM sailkatze-algoritmoak darabiltzate subjektibotasuna sailkatzeko, % 92 eta % 90eko doitasunak lortzen dituzte, hurrenez hurren.

(Wang eta Fu, 2010) lanak sentimenduen dentsitatean oinarritutako Naive Bayes sailkatzaile bat proposatzen du, txineraren subjektibotasuna detektatzeko. “Conditional Random Field“ eredu estatistikoak ere proposatu dira subjektibotasuna detektatzeko atazan (Das eta Bandyopadhyay, 2009a).

Edozein modutan, bi kasuetan polaritate-lexikoak jakintza-baliabide garrantzitsuak dira. Ikasketa automatikoko sistemari lexikoen ezagutza gehitzeak emaitzak hobetzen ditu (Yu eta Hatzivassiloglou, 2003). Chaovalit eta Zhou (2005) ikertzaileek egindako esperimentuetan, ikasketa automatikoak erregeletan oinarritutako metodoa gainditzen du, polaritatea hautemateko momentuan.

Dena dela, euskara bezalako hizkuntzetan, entrenatzeko datu-kopurua oso urria da. Dokumentu-mailan sailkatutako datu-sortak eraikitzeko, maiz metodo automatikoak erabiltzen dira literaturan. Pelikula (Turney, 2002) nahiz produktuen kritikak biltzen dira (Pang et al., 2002), eta erabiltzaileak berak emandako puntuazioen arabera sailkatu.

Maks eta Vossen (2012) ikertzaileen lanean, albisteak corpus subjektibo gisa biltzen dira, eta Wikipediako artikuluekin alderatu, azken horiek objektibotzat jotzen dira eta.

Eskuz sortutako baliabideak ere badira, hala nola blogetako testuen JDPA corpusa (Kessler et al., 2010) eta MPQA albisteen corpusa (Wiebe et al., 2005). Subjektibotasun-sailkapena duten esaldi-sortak etiketatzea oso lan nekeza da. Etiketatzea fidagarria izan dadin, aditu batek baino gehiago hartu behar du parte, eta prozesu horretan, desadostasunak sortzen dira (Wiebe eta Mihalcea, 2006).

Ondorioz, esaldi-kopurua ere txikiagoa izaten da. Estrategia bat da esaldiak subjektibotzat edo objektibotzat hartzea, barnean esaera edo hitz subjektiboak dituzten ala ez aztertuta (Yu eta Kübler, 2011).

Subjektibotasun-lexikoei dagokienez, dagoeneko lehen aipatu dugun bezala, ikerketa eta baliabide gehienak ingeleserako landu dira. Gaur egun, OpinionFinder (Wilson et al., 2005a) eta SentiWordNet (Esuli eta Sebastiani, 2006) lexikoak erabiltzea oso ohikoa da sentimenduen analisian.

Alabaina, baliabide horiek ez dituzte baliabide urriko hizkuntzak kontuan hartzen. Proposatutako teknikak hizkuntzarekiko independenteak dira hein batean; baina, erabili ahal izateko, zenbait baliabide sortzea beharrezkoa da, eta batez ere polaritate-lexikoak edo subjektibotasuna/polaritatea etiketatuta duten corpusak, ikasketa automatikoko sistemak entrenatzeko.

Baliabide horiek eskuratzeko, bi aukera daude: horiek hutsetik sortzea ala existitzen direnak hizkuntza berrira egokitzea. Eskuz sortzeak kostu handia du, eta hizkuntza askoren kasuan, ez da errentagarria. Horrenbestez, metodo automatikoetara, edo, gutxienez, erdi-automatikoetara jotzea besterik ez zaigu geratzen.

Subjektibotasun-lexikoak hutsetik sortzeko, corpusetara jotzen da, eta corpusetatik hitz subjektiboak erauzten dira, metodo estatistikoen bitartez (Yu eta Hatzivassiloglou, 2003; Kaji eta Kitsuregawa, 2007; Maks eta Vossen, 2012). Estrategia bat da hazi-hitz batzuetatik abiatuta hitz horiekiko antzekotasun distribuzionalik handiena duten hitzak bilatzea (Turney, 2002; Kanayama eta Nasukawa, 2006).

Beste aukera bat da subjektibo gisa etiketatuta dauden testuak objektibo gisa dauden testuekin alderatzea, lexiko dibergentea erauziz (Rayson eta Garside, 2000; Kilgarriff, 2001; Maks eta Vossen, 2012). Ideia hori aurrera eramateko, sailkatutako testuak behar dira aurrena.

Literaturan erabilitako beste estrategia bat da hasierako hitz zerrenda bat handitzea, sinonimoak edo tesauro batean lotutako hitzak erabiliz (Hu eta Liu, 2004; Kim eta Hovy, 2004; Esuli eta Sebastiani, 2005). Estrategia horren arazoa da WordNet edo horren gisako baliabide gehigarriak eskatzen dituela.

(Mihalcea et al., 2007) eta (Banea et al., 2008) lanek ingelesezko baliabideak errumanierara itzultzen dituzte, hiztegien eta itzulpen automatikoko tekniken bidez. Perez-Rosasen

ikertzaile-taldeak (Perez-Rosas et al., 2012) SentiWordNet eta WordNet baliatzen ditu ingelesetik gaztelaniara lexikoa itzultzeko. Itzulpen-prozesuak, hala ere, baliabideen kalitatea murrizten du, (Mihalcea et al., 2007) eta galera hori erregeletan oinarritutako sistemetan handiagoa edo txikiagoa izan daiteke, ikerketaren arabera.

Mihalcearen taldeak, bere aldetik, 2007an lan bereziki esanguratsua egin zuen hizkuntza arteko proiekzioak erabiliz egindako sentimendu-analisi eleanitzaren gainean. Lan horretan, zubi-baliabideak eta corpus paraleloak erabili zituzten, helburuko hizkuntzarako (errumaniera) esaldi-mailako subjektibotasun-sailkatzaileak sortzeko.

Emaitzetako bat bereziki interesgarria izan zen: lexikoko sarreren zati txiki batek bakarrik mantentzen ditu polaritateak itzulpena egin ondoren, lexikoa itzultzean aurkitutako polaritatea eta subjektibotasuna galtzen delako. Corpusean oinarritutako itzulpenak, aldiz, hitzen anbiguetatea hobeto mantentzen dute, eta subjektibotasuna mantendu egiten da esaldien itzulpenetan.

Banea et al. (2011) ikertzaile-taldeak, subjektibotasuna sailkatzeko metodoak aztertu zituen errumanierarako, eta horien artean, lexikoa hutsetik sortzea edo itzultzea egokiagoa den. Corpusetan oinarritutako lexikoak emaitzak hobeak izan zituen. Itzulitako lexikoak hainbat itzulpen oker zituela konturatu ziren, itzulpen-prozesuan sortutako anbiguotasunak direla eta. Ikerketa hura albisteez osatutako corpus baten gainean burutu zen.

3 Metodologia

Hurrengo lerroetan gure ikerketa aurrera eramateko erabili diren oinarrizko baliabideak deskribatzen dira. Baliabide horietako batzuk, dokumentu- eta esaldi-bildumak kasu, bereziki lan honetarako sortuak izan dira, eta horien eraikitze prozesua ere azaltzen da. Horietaz gain, guk sortutako subjektibotasun-lexikoak eraikitzeko erabilitako tresnak aurkezten dira, hau da, erabilitako baliabide lexikalak, tresna linguistikoak edo neurri estatistikoak. Haatik, lexikoen eraikitze-prozesuari buruzko xehetasunak 4 atalean ematen dira, gure esperimentuen parte direlako.

3.1 Datu-sortak

Aurreko atalean aipatu dugunez, subjektibotasunaren detekzioa burutzeko, subjektibotasuna etiketatuta duten corpusen edo datu-sorten behar dira. Subjektibotasuna detektatzeko sistema bat garatzea helburu izanik, arazo nagusietako bat erreferentziazko datu-sorta bat lortzea da.

Lan honetako esperimentuak aurrera eramateko, subjektibotasuna dokumentu-mailan etiketatuta duen corpus bat beharko genuke gutxienez, hau da, dokumentu bakoitza objektiboa edo iritzizkoa den markatuta duen corpusa. Ikasketa automatikorik aplikatuko

ez badugu ere, erregeletan oinarrituta dagoen sistema hau ebaluatzeko ezinbestekoa dugu etiketatutako corpusa.

Horien artean nagusiena, sistemaren garapena eta optimizazioa aurrera eramateko bildutako corpusa da. Alor periodistikoko corpus bat osatu dugu, Berria[4] egunkariko berriak bilduta. Corpusak 2006-2009 urteetako albisteak biltzen ditu (37 miloi hitz), eta egunkariko atalaren arabera sailkatuta daude.

Berria egunkariaren web orrietatik datuak lortzeko, crawling sistema bat garatu dugu. Behin dokumentuak lortutakoan, wrapper bat erabili da bertatik testu garbia eta dokumentuari dagozkion metadatuak eraucteko. Gure hurrengo arazoa albiste subjektiboak eta objektiboak sailkatzea izan da. Sailkapen hori modu automatikoan egin dugu.

Lehen esan dugu legez, albiste bakoitzaren atalari buruzko informazioa ere eskuratu dugu. Informazio hori baliatuz, "Iritzia" izeneko atalean sailkatutako albiste guztiak gure corpus subjektiboan sartu ditugu. Albiste objektibo gisa, ordea, "Harian" ataleko berriak hartu dira. Atal horrek gertakarien inguruko berriak biltzen ditu, eta printzipioz, informazio objektiboa jasotzen dute berriek.

Argudia liteke objektiboen artean gaizki sailkatutako artikuluak daudela. Izan ere, albisteak orokorrean objektibotzat hartzen badira ere, Wiebe et al. (2001) egileek Wall Street Journal egunkariko corpus baten gainean egindako azterketak objektibotzat hartutako dokumentuetan esaldien % 44 subjektiboak zirela.

Tamalez, metodologia hori zen gure baliabideen mugek corpus etiketatu bat lortzeko ahalbidetzen diguten aukerarik onena. 1 taularen lehen lerroan, corpusaren tamainaren inguruko datuak ikus daitezke.

3.1.1 Dokumentu-mailako datu-sortak

Corpus horretatik bi datu-sorta osatu ditugu dokumentu-mailan lan egiteko. Bat sistemaren garapenerako erabili dugu, eta $Berria_{Train}$ izena eman diogu. Corpus horrek 21.320 dokumentu ditu (ikus 1 Taula), % 50 objektiboak eta % 50 subjektiboak, alegia. Corpus hori osatzeko, bi gauza eduki dira kontutan: batetik, dokumentu objektiboen eta subjektiboen kopurua orekatua izatea, eta, bestetik, garapenerako eta ebaluaziorako banaketa egoki bat egitea.

1 taulan ikus dezakegunez, hasierako corpusean dokumentu subjektiboen kopurua askoz txikiagoa da, objektiboena baino. Ondorengo bide hau hartu dugu: dokumentu subjektiboen artean, % 80 ausaz aukeratu dugu, garapenerako, eta beste % 20, berriz, ebaluaziorako, hots, 10.661 eta 2.665, hurrenez hurren.

Kopuruetan oreka mantenduz, berri objektiboen artetik beste horrenbeste albiste hautatu dira, ausaz kasu honetan ere. Horrela osatu ditugu $Berria_{Train}$ eta $Berria_{Test}$ deitu

[4] http://berria.info

ditugun datu-sortak.

Corpus	**#hitz**	**#dok**	**#hitz subj.**	**#dok subj.**	**#hitz obj.**	**#dok obj.**
Berria	20.402.121	75.892	3.810.857	13.325	16.591.264	62.567
$Berria_{Train}$	6.962.081	21.322	3.645.059	10.661	3.317.022	10.661
$Berria_{Test}$	1.718.337	5.330	888.597	2.665	829.740	2.665

Taula 1: Berriako corpusaren eta bertatik ateratako datu-sorten estatistikak.

Dokumentu-mailako subjektibotasunaren detekzioa ebaluatzeko, bigarren corpus bat ere prestatu dugu, lehen corpusaren metodologia berdina jarraituz. Kasu honetan, Gara[5] egunkariko albisteak bildu ditugu. Oraingoan ere, iritzien atalean sailkatutako albisteak subjektibotzat hartu dira. Albiste objektiboen zerrenda osatzeko, berriz, Euskal Herriko eta mundu-mailako albisteak dituzten atalak hartu ditugu.

2 taulan ditugu corpusari eta ebaluatzeko erabilitako $Gara_{Test}$ datu-sortari dagozkien estatistikak. Kasu honetan, albiste subjektibo guztiak sartu ditugu ebaluaziorako datu-sortan, eta objektiboak ausaz aukeratu dira, subjektiboen kopuru berera iritsi arte.

Corpus	**#hitz**	**#dok**	**#hitz subj.**	**#dok subj.**	**#hitz obj.**	**#dok obj.**
Gara	3.435.769	10.312	1.267.531	4.669	2.168.238	5.643
$Gara_{Test}$	3.340.960	9.338	1.267.531	4.669	2.073.429	4.669

Taula 2: Garako corpusaren eta bertatik ateratako datu-sortaren estatistikak.

3.1.2 Esaldi-mailako datu-sortak

Subjektibotasuna detektatzeko darabilgun metodologiak iritzizko esaldiak detektatzeko balio duen ere aztertu dugu lan horretan. Azterketa hori aurrera eramateko, subjektibo eta objektibo gisa sailkatutako esaldi-sorta bat osatu dugu.

Sorta hori eratzeko, $Berria_{Test}$ corpusetik 400 esaldi erauzi ziren ausaz, 200 dokumentu objektiboren artetik eta 400 dokumentu subjektiboren artetik. Esaldi horiek pertsona bakarrak sailkatu zituen 6 orduko lanean. Honako sailkapen-irizpide hau erabili genuen: subjektibotzat jo ziren erabat objektiboak ez ziren esaldiak. Ikus dezagun hurrengo adibidea:

- *"Giza katea egin dute 10:00etan Baionan , Euskal Herriko Laborantza Ganberak Frantziako Administraziotik jasandako erasoak salatzeko."*

Esaldiaren lehen zatia objektiboa da ("giza katea egin dute"), baina bigarren zatiak egoera pertsonal bat adierazten du ("jasandako erasoak"eta "salaketa"). Horrelako kasuak subjektibotzat jo ziren. 3.1.2 taulak etiketatzearen emaitzak erakusten ditu.

[5]http://gara.net

Batetik, harritzekoa da esaldi subjektibo gisa sailkatutako perpausen kopurua baxua dela orokorrean (% 21), eta, are gehiago, aintzat hartzen badugu sailkapen irizpidea subjektiboen aldekoa zela.

Bestalde, esaldi subjektiboen kopurua askoz altuagoa da dokumentu subjektiboen taldean (% 37) dokumentu objektiboen taldean baino (% 5). Esaldi horiek gure corpusaren lagintzat hartuko bagenitu, zenbaki horiek corpusa modu egokian sortuta dagoenaren aztarna litzateke.

	Dok. Subj. (200)		**Dok. Obj (200)**	
	Subj.	**Obj.**	**Subj.**	**Obj.**
Eskuzko sailkapena	74	126	11	189

Taula 3: Esaldien estatistika

Ebaluaziorako esaldi-sorta osatzeko subjektibo gisa sailkatutako 85 esaldiak hartu ziren, eta beste horrenbeste ausaz aukeratu ziren objektibo gisa sailkatutako esaldien artetik. Beraz, 170 esaldiko sorta osatu dugu esaldi-mailako subjektibotasunaren ebaluaziorako.

3.2 Subjektibotasun-Lexikoak

Aurreko ataletan aipatu dugun bezala, subjektibotasun lexikoak testuen subjektibotasunaren detekziorako baliabiderik garrantzitsuenetako bat dira. Lexiko horiek sortzerakoan, kostu baxuko bi bide aztertu ditugu lan honetan.

Alde batetik, beste hizkuntza batean existitzen diren lexikoak euskarara itzultzea edo proiektatzea, hiztegi elebidunen bitartez, eta, bestetik, corpusetatik subjektibotasuna adierazten duten terminoak automatikoki erauztea, teknika estatistikoak erabilita.

Lehen bideak itzulpen prozesuaren ondorioz ematen den kalitate-galerari aurre egin behar dio; baina, bestalde, baliabide independentea da printzipioz, eta edozein datusortaren gainean erabil daiteke. Corpusetan oinarritutako lexikoak, aldiz, garatutako corpusaren menpekoak dira, eta, ondorioz, beste domeinu batzuetan aplikatzean errendimendugalera bat eragin lezake.

Proiekzioari dagokionez, ingeleserako sortu zen OpinionFinder Wilson et al. (2005a) lexikoa itzuli dugu. Lexiko horretan, hitz subjektiboak besterik ez daude, eta hitzen hsubjektibotasuna bi mailatan sailkatzen du: altua ala baxua. 4 taulak sailkapen horren banaketa ematen du.

	#subj. altua	**#subj. baxua**	**#sarrera**
OpinionFinder	4.743	2.188	6.931

Taula 4: OpinionFinder hiztegiaren hitzen sailkapena

Lexikoaren itzulpena burutzeko, Elhuyar Fundazioaren euskara-ingelesa[6] hiztegia erabili dugu. 5 taulan, hiztegiaren ezaugarriak ikus daitezke.

Dictionary	**#entries**	**#pairs**	**ambiguity level**
$D_{en \rightarrow eu}$	17,672	43,021	2.43

Taula 5: Ingelesezko subjektibotasun-lexikoa itzultzeko erabilitako hiztegi elebidunaren estatistikak.

Subjektibotasun-lexikoak corpusetatik automatikoki erauzteko, elkartze-neurrietan oinarritutako teknika estatistikoak erabili dira. Testu subjektiboen eta testu objektiboen artean dagoen lexiko dibergentea erauztea da gure helburua.

Hurbilpen horren oinarrian dagoen helburua da antzematea zeintzuk diren dokumentu subjektiboetan askotan agertzen diren hitzak eta dokumentu objektiboetan gutxitan agertzen direnak, hau da, dokumentu subjektiboekin elkartze-mailarik altuena duten hitzak erauztea da gure asmoa.

4.2 atalean lexiko horiek erauzteko burutu diren esperimentuak modu zehatzagoan deskribatzen dira. Corpus bateko hitzek beste corpus batekin duten elkartze-maila kalkulatzeko, hainbat neurri estatistiko erabiltzen dira. Hurrengo azpiatalean, lan honetan erabilitakoen azalpen laburra eskaintzen dugu.

3.2.1 Elkartze-neurriak

subjektibotasun-lexikoak lortzerakoan, gure metodoa honako hausnarketa honetatik abiatuta da: zein dira testu subjektiboen hitzik esanguratsuenak, testu objektiboekin alderatuz? Oro har, elkartze-neurriak erabiltzen dira corpus batean hitz-segida jakinak elkarrekin agertzeko zer nolako probabilitatea edo joera duten neurtzeko (Daille, 1995; Evert, 2005).

Aitzitik, testu edo corpus bateko hitzik esanguratsuenak zein diren topatzeko ere erabili ohi dira, corpusen arteko "diferentzia" neurtzeko, alegia (Kilgarriff, 2001; Rayson eta Garside, 2000). Hitz batek bi corpusetan duen garrantzia 6 gertakizun-taularen bidez adierazten da.

Bertan, a eta b hitzaren maiztasunak dira, $Corp_1$ eta $Corp_2$ corpusetan, hurrenez hurren, eta c eta d balioek corpusetako gainontzeko hitz guztien maiztasunen batura adierazten dute. $np1$ eta $np2$ balioek bi corpusen tamaina adierazten dute, hitzetan, eta $n1p$ eta $n2p$ w hitzaren maiztasuna populazio osoaren barnean.

Jarraian, erabiliko ditugun elkartze-neurriak zerrendatzen dira, haien formula 6 taulako balioen arabera ematen delarik:

[6] http://www.elhuyar.org/hizkuntza-zerbitzuak/EU/Hiztegi-kontsulta

	$Corpus_1$	$Corpus_2$	
w	a	b	$a + b = n1p$
not w	c	d	$c + d = n2p$
	$a + c = np1$	$b + d = np2$	$a + b + c + d = npp$

Taula 6: Hitz baten gertakizun-taula.

Pointwise Mutual Information (PMI): Pointwise Mutual Information (PMI) neurriak hitz batek agertzeko izango lukeen probabilitatearen ($np1 * n1p/npp$) arteko desbideratzea neurtzen du, kontuan izanik hitz hori $Corp_1$ corpusean duen maiztasuna $n11$ eta hitz horrek bi corpusetan dituen agerpen-kopurua erabat independenteak direla.

$$PMI(w, corp_1, corp_2) = \log(a/(np1 * n1p/npp))$$

Neurri horrek maiztasun baxua duten hitzei esangura-maila altua esleitzeko joera du.

(Log) Odds ratioa: Odds ratioak hitz bat bi corpusetan ala batean ere ez agertzearen eta bi corpusetako batean bakarrik agertzearen arteko ratioa kalkulatzen du.

$$Odds(w, corp_1, corp_2) = a * d/c * b$$

Neurri horrek corpusen tamainarekiko mendekotasun handia du, eta tamaina txikiko eta ertaineko corpusekin lortzen diren balioek desbiderapenak izaten dituzte. Efektu hori murrizteko asmoz, neurriaren beste aldaera bat erabiltzen da, log-odds ratio izenekoa.

$$log - odds(w, corp_1, corp_2) = \log((a + 0,5) * (d + 0,5)/(c + 0,5) * (b + 0,5))$$

Fisher: Fisher elkartze-neurria Fisherren test zehatzetik dator. Fisherren testak w hitz batek egun bi corpusetan zer banaketa duen jakiteko probabilitate zehatza ematen du, kontuan izanik w hitza bi corpusetan agertzeko probabilitatea berdina dela.

$$Fisher = \frac{\binom{n1p}{a}\binom{n2p}{c}}{\binom{npp}{np1}} = \frac{(n1p)!(n2p)!(np1)!(n2p)!}{a!b!c!d!npp!}$$

Neurri honek probabilitate zehatza kalkulatzen du, estimazioetan oinarritu gabe. Alabaina, horrek kostu handia dakar konputazionalki. Kostu hori murrizteko, alde bateko Fisher testak erabiltzen dira. Guk ezkerreko Fisher testa erabiliko dugu, eta ezkerreko testak corpus subjektiboan hitzak egun duen banaketan baino gutxiagotan agertzeko probabilitatea ematen du. Horrela, probabilitate altu batek adierazten du hitza egungo corpus

subjektiboan baino gutxiago agertzea zaila dela, eta, hortaz, subjektibotasun-maila altua duela.

Ezkerreko Fisher testa kalkulatzeko, $n11$ balioa baino txikiagoa edo berdina duten probabilitateak bakarrik hartzen dira kontuan goiko formulan.

Egiantz-arrazoia (LLR): LLR neurriak w hitz baten agerpenaren "harrigarritasuna" neurtzen du. w hitzak egungo populazioan ($corp_1$ eta $corp_2$) dituen agerpenen eta populazio-eredu hipotetiko batean izango lituzkeen agerpenen arteko dibergentzia kalkulatzen da. Gure kasuan, populazio-eredu hipotetikoan hitz guztiek ausaz agertzeko probabilitate bera dute. Horrenbestez, dibergentzia-balio hori zenbat eta handiagoa izan, orduan eta arrigarrikoagoa da w hitzak agertzeak ez dela ausazko kontua, eta, ondorioz, hitz subjetiboak direla (beste era batera esanda, w hitzaren agerpena $corp_1$ corpusean esanguratsua dela $corp_2$ corpusarekin alderatuta).

$$LLR(w, corp_1, corp_2) = 2 \sum_{i,j} k_{ij} \log \frac{n_{ij}}{m_{ij}}.$$

3.3 Etiketatze linguistikoa

Gure ikerketa aurrera eramateko, ezinbestekoa zaigu bildutako corpusak linguistikoki prozesatzea.

Batetik, itzulitako lexikoa erabili ahal izateko, hitzei dagozkien lemen informazioa behar dugu. Automatikoki erauzitako hiztegien kasuan, formekin lan egin genezake; baina horrek neurri estatistikoetan eragina izango luke, datuen dispertsioa dela eta.

Bestetik, hitzen kategoria gramatikala ere jakin nahi dugu, ezaugarri horrek detekzioan duen eragina aztertzeko. Gure corpusak informazio horrekin etiketatzeko, IXA taldeak garatutako Eustagger (Aduriz eta de Ilarraza, 2003) tresna erabiliko dugu.

4 Gure hurbilpena

Atal honetan, subjektibotasuna detektatzeko eraiki dugun sistema deskribatuko dugu. Erabilitako oinarrizko baliabideak aurreko atalean azaldu badira ere, lan honen esperimentuek barne hartzen dute baliabide horietako batzuk eraikitzea; hori da, hain zuzen ere, sortutako subjektibotasun-lexiko ezberdinen kasua. Hurrengo lerroetan, prozesu horien zehaztasunak ematen dira.

HAP masterra

Bestalde, atal honetan sistemaren errendimendua ebaluatzeko diseinatu diren esperimentuak deskribatu ditugu, baita sortutako baliabideek sistemaren portaeran duten eragina ere.

4.1 Subjektibotasuna detektatzeko algoritmoa

Subjektibotasunaren detekzioa burutzeko, erregeletan oinarritutako sailkatzailea inplementatu dugu. t testu batek duen subjektibotasun-maila kalkulatzeko, sailkatzaileak L_{eu} subjektibotasun-lexikoaren informazioa hartzen du oinarritzat. Testu osoaren subjektibotasun-maila testuko hitzen subjektibotasun pisuen batazbestekoa da, 1 ekuazioak adierazten duen moduan.

$$sub(t) = \sum_{i,j} bal(w)/\#w \tag{1}$$

non $bal(w) = L_{eu}(w)$ w hitzak subjektibotasun-lexikoan duen pisua den eta $\#w$ testuaren hitz kopurua den.

Corpusetatik erauzitako lexikoen kasuan L_{corp}, $bal(w)$ funtzioak w hitzak corpus objektiboetan eta subjektiboetan dituen maiztasunen arteko dibergentzia adierazten du. Dibergentzia hori subjektibotasun-lexikoan dago erregistratuta, eta lexikoa sortzeko erabili den elkartze-neurriaren arabera aldatzen da (ikus 3.2.1 atala). Corpus subjektiboaren aldeko dibergentziak subjetibitate-maila altua adierazten du, hau da, balio altu batek subjektibotasun maila altua adierazten du.

$sub(t)$ formularen balioa, hortaz, bi faktorek baldintzatzen dute: subjektibotasun-mailak eta hitz-kopuruak. Hitz subjektibo gutxi batzuk besterik ez dituen testu batek balio altua lor lezake, hitz horiek subjektibotasun maila oso altua izango balute. Hori ekiditeko, subjektibotasuna testuaren hitz-kopuruarekin zatitzen da, batazbesteko balioa lortuz.

Ingelesetik itzulitako L_{Trans} lexikoak ez du subjektibotasun-mailaren gaineko informaziorik. Arazo horri erantzuteko, hitzaren presentzia erabiliko dugu $bal(w)$ funtzioaren balio gisa; hots, w hitza lexikoan badago, 1 balioa emango diogu, eta, 0, bestela:

$$bal(w) = \begin{cases} 1 & \text{if } w \in L_{Trans}, \\ 0 & \text{else} \end{cases} \tag{2}$$

$Sub(t)$ kalkulatu eta gero, α atalase bat gainditzen duten testu-unitateak subjektibo gisa sailkatzen dira, eta gainontzeko unitateak, ordea, objektibotzat. Atalase horren balio optimo bat topatzeko, $Berria_{Train}$ garapen-corpuseko testuen subjektibotasun-balioak kalkulatzen dira, eta asmatze-tasa maximizatzen duen atalasea topatu.

HAP masterra

4.2 Subjektibotasun-lexikoak

Subjektibotasun-lexikoak erauzteko jarraitu diren bi bideak aipatu ditugu dagoeneko, hau da, beste hizkuntza bateko lexikoaren itzulpena edo proiekzioa eta corpusetatik lexikoa automatikoki erauzteko metodoa. Ondorengo paragrafoetan, bi hurbilpen horien inguruko xehetasunak emango ditugu, baita lexikoak eraikitzeko jarraitutako prozesuen deskribapen zehatza egin eta lortutako lexikoen inguruko datu zehatzak eman ere.

Lexiko horiek aurreko atalean azaldutako algoritmoaren baliabide nagusia osatzen dute. Lexiko ezberdinek ataza bakoitzean duten portaera ebaluatu da; baina atal honetan, bi hurbilpenen arteko konparaketa nola egin den azalduko dugu (4.2.3. Lexikoen ezaugarriak ezberdinak direnez, corpusetan oinarritutako lexikoak egokitu egin ditugu, ebaluazioa bidezkoa izan dadin.

4.2.1 Ingelesetik itzulitako lexikoa

atalean azaldu dugun legez, OpinionFinder Wilson et al. (2005a) subjektibotasun-lexikoa itzuli dugu, Elhuyar Fundazioaren ingelesa-euskara hiztegi elebiduna erabiliz. 5 taulan ikus daitekeenez, itzulpen prozesu horretan anbiguotasunak sortzen dira, sarrera bakoitzak bi itzulpen edo gehiago dituelako, bataz beste.

Anbiguotasun horiek modu automatikoan itzultzeko aukera dago (Saralegi eta Lacalle, 2009). Lan honetarako, lehen itzulpena aukeratuko dugu itzulpen zuzen gisa (Mihalcea et al., 2007).

Itzulpen-prozesuaren amaieran lortutako lexikoak jatorrizkoaren sarreren erdia baino gutxiago ditu (ikus 7 taula). Ondorioz, aurreikusi genezake, estaldura baxuagoa izanik, testuaren subjektibotasun-maila kalkulatzeko informazio gutxiago erabiliko dela, eta, horrenbestez, subjektibotasun hori kalkulatzerakoan arazoak sor litezkeela.

	OpinionFinder (en)	L_{Trans} **itzulitako Lexikoa**
Subj. altuko hitzak	4.743	1.519
Subj. baxuko hitzak	2.188	1.169
Osotara	6.931	2.788

Taula 7: Itzulitako lexikoari dagozkion estatistikak.

4.2.2 Automatikoki erauzitako lexikoak

Lexikoak sortzeko bigarren estrategia corpusetatik lexiko subjektiboa erauztean datza. Guk jarraitutako estrategia izan da subjektibotasun-maila desberdina duten bi testu-bildumetatik lexiko subjektiboa erauztea Maks eta Vossen (2012). Horretarako, $Berria_{Train}$

garapenerako corpusa bi zatitan banatu dugu: testu subjektiboak batetik ($Berria_{Train}SUBJ$), eta objektiboak bestetik, ($Berria_{Train}OBJ$).

Berria corpus osoa erabil genezake lexikoa sortzeko, corpus handiagoa erabiltzeak lexikoa osatzeko informazio aberatsagoa emango ligukeelako. Alabaina, lexiko horrek estaldura zabalagoa lortuko luke testeko dokumentuen gainean, eta abantailazko egoera batean legoke, horiek sailkatzeko garaian. Modu honetan, ebaluaziorako datuak garapeneko datuetatik ahalik eta independenteenak izan daitezen bermatu nahi dugu.

Ondoren, $Berria_{Train}SUBJ$ eta $Berria_{Train}OBJ$ zatien arteko terminologia alderatu da hainbat neurri estatistikoren arabera, lexiko dibergentearen bila. Zehazki, corpus subjektiboaren adierazgarri diren hitzak topatu nahi ditugu, hau da: zer hitz diren esanguratsuen testu subjektiboetan, testu objektiboekin alderatuz gero (Kilgarriff, 2001).

Bi corpusak Eustaggerrekin linguistikoki etiketatu dira, eta lemekin lan egin dugu. Corpus subjektiboko lema guztien elkartze-maila kalkulatu dugu, "erreferentziazko" corpus objektiboarekin alderatuta. Elkartze-maila hori 3.2.1 ataleko neurrien arabera kalkulatzen da.

Errore ortografikoak, hitz arraroak eta beste hizkuntzatako hitzak lexikoan sar daitezen ekiditeko, hitz guztiek corpus subjektiboan gutxienez 2 agerpen izan behar zituzten nahitaez, lexikoan sartu ahal izateko ($\forall w \in L_i,\ freq(w, Berria_{Train}SUBJ) > 1$).

L_{LL}		L_{Fisher}		L_{PMI}		L_{ODDS}		$L_{Log-Odds}$	
gu	10764.98	eta	1.00000011590148	zurito	2.4011	amagoi	368.28	lan-indar	6.56
ni	7040.96	ukan	1.00000006535656	berenjena	2.4011	lan-indar	351.15	halaz	6.07
ez	6439.96	ez	1.00000006416492	arabiar-palestinar	2.4011	zerbeza	231.24	habermas	5.97
hau	5714.33	ere	1.00000001445008	belauntzi	2.4011	praileaitz	218.4	ostraka	5.82
edo	4500.87	gu	1.0000000055863	produkzio-sistema	2.4011	halaz	214.11	halatan	5.82
euskaldun	3991.85	*ezan	1.00000000414656	testu-liburu	2.4011	habermas	192.70	upv-ehu	5.72
bezala	3568.97	oso	1.000000004114	kilikagarri	2.4011	onkeria	184.14	naski	5.69
jakin	3523.27	baino	1.00000000198889	berpasatu	2.4011	ostraka	167.01	adornu	5.69
omen	3329.00	hau	1.00000000190684	maiseaketa	2.4011	halatan	167.01	aktibo-prezio	5.63
euskara	3188.54	gabe	1.00000000177986	klaudio	2.4011	upv-ehu	149.88	fredi	5.63
...		...		...		...		...	

Taula 8: Corpusetatik erauzitako subjektibotasun-lexikoen lehen 10 sarrerak, neurri bakoitzaren arabera.

Gure subjektibotasun-lexikoak elkartze-neurriek emandako pisuen arabera sailkatuta dituzte hitzak: zenbat eta subjektiboagoak izan, orduan eta gorago daude. Elkartze-neurri bakoitzak balio-eskala bat du, 4.2.2 taulako hiztegien adibideek erakusten duten bezala.

Horren aurrean, erabaki beharko genuke subjektibotasun-maila adierazgarria duten hitzak zein baliotaraino heltzen diren neurri bakoitzean, edo, bestela, lexiko guztien kasuan b hitz-kopuru bat zehaztu eta pisurik handiena duten b hitzak hartu.

Guk, horren ordez, lexikoetan hitz guztiak uztea erabaki dugu. Nahiz eta subjektiboak ez diren hitz batzuk lexikoan sartu, pisuek hitz horien eragina mugatzen dute.

4.2.3 Itzulpen- eta erauzketa-estrategien ebaluazioa

Lexikoa sortzeko bi hurbilpenen arteko ebaluazioa egiteko, $Berria_{Test}$ eta $Gara_{Test}$ datusortetako dokumentuak sailkatzeko ataza diseinatu dugu. Iturri ezberdinetako bi bilduma ebaluatzeak ondorio sendoagoak atertzeko aukera ematen digu. Gainera, bi hurbilpenen arteko konparaketa bidezkoa izan dadin, neurriak hartu ditugu.

L_{Trans} itzulitako lexikoak termino bakoitzaren subjektibotasun-pisuari buruzko informaziorik ez du. Ondorioz, $sub(t)$ funtzioaren balioa kalkulatzerakoan hitz guztiek pisu berdina dute. Corpusetik erauzitako lexikoek, informazio hori erabiltzeko gaitasuna izanik, subjektibotasun-balio zehatzagoak lortzeko aukera dute, eta horrek L_{Trans} lexikoaren aurrean abantaila-egoera batean jartzen ditu. Gure ebaluazioan, abantaila-egoerarik ematen ez dela ziurtatu nahi izan dugu, eta helburu horrekin, corpusetatik erauzitako lexikoak egokitu ditugu.

Esperimentu horietarako, LLR neurriarekin sortutako lexikoa bakarrik erabili dugu, eta lexiko horren bi aldaera sortu ditugu. Bi lexiko berriak LLR neurriarekin lortutako lexikoan oinarrituta daude, baina ez dute subjektibotasun-pisuen baliorik. L_{LL} lexikoak gutxieneko maiztasuna gainditzen duten lema guztiak biltzen ditu, eta, hortaz, subjektibotasun-pisurik gabe erabiltzeko, hitz subjektiboak direnak bakarrik hartu beharko ditugu.

Horretarako, lexiko-sarrera kopuru zehatz bat finkatzea erabaki dugu. Lehenengo aldaerak (LL_PRES5000) LLR neurriarekin lortutako lexikotik pisurik altuena duten lehen 5.000 sarrerak baino ez ditu. L_{Trans}-ekin alderatuz, lexiko horren tamaina handia da, eta, horrenbestez, estalduraren diferentzia zeharo nabarmena da.

Horregatik, bigarren aldaera bat sortu dugu, LL_PRES3000 deitu duguna. Aldaera horrek LLR neurriarekin lortutako lexikoko lehen 3.000 sarrerak baino ez ditu. Lexiko horrek L_{Trans} lexikoaren sarrera-kopuru antzekoa du, 9 taulan ikus daitekeenez.

Lexikoa	**Sarrera kopurua**
$L_{LL\_PRES5000}$	5.000
$L_{LL\_PRES3000}$	3.000
L_{Trans}	2.788

Taula 9: Presentzia bakarrik kontutan hartzen duten lexikoen tamaina.

Subjektibotasun-mailaren informaziorik ez dagoenez, dokumentuen subjektibotasuna kalkulatzerakoan hitzen agerpena bakarrik hartzen da kontuan, hots, t testuaren $sub(t)$ kalkulatzerakoan 2 ekuazioak deskribatutako aldaera erabiliko da.

4.3 Dokumentu-mailako vs. esaldi-mailako subjektibotasuna

Lan honetan dokumentu-mailako subjektibotasuna landu dugu nagusiki. Dena den, azken urteetan lan gehienak esaldi-mailako subjektibotasuna detektatzera bideratu dira. Gure

etorkizuneko bidea ere hori dela uste dugu, bereziki Twitter-en moduko baliabideetatik iritzia erauzteko gai izan nahi badugu, bertan mezuak esaldi bakarrera mugatzen baitira gehienetan.

Dagoeneko esan dugun legez, zoritxarrez oso lan nekagarria da subjektibotasun-sailkapena duten esaldi-sortak sortzea. Horren erakusgarri dugu 3.1.2 atalean azaldutako esaldi-sorta etiketatzeko 6 orduko lana behar izana (eta pertsona bakarrak egin zuen). Hasierako esaldietatik, 400 esaldietatik 170 esaldi lortu genituen. Sorta hori sistema baten garapenerako txikiegia den arren, ebaluaziorako nahikoa izan liteke.

Gure sistemak testu laburragoekin zer nolako errendimendua duen aztertu nahi izan dugu. Horretarako, gure algoritmoa inolako aldaketarik gabe aplikatu dugu. Esaldiak sailkatu dira, eta dokumentu-mailan lortutako emaitzekin alderatu ditugu.

Gure algoritmoaren oinarriari erreparatuta, agerikoa da sarrerako testu-unitatearen luzerarekiko mendekotasunik ez du, $sub(t)$ funtzioaren balioa testuaren luzerarekin normalizatuta baitago. Haatik, sistema dokumentu luzeagoak sailkatzeko optimizatuta dago, eta, ondorioz, errendimendua okertzea ulergarria litzateke.

4.4 Berria vs. Gara

Gure sistema optimizatzeko, $Berria_{Train}$ corpusa erabili dugu, baita subjektibotasun-lexikoak eraikitzeko ere. Nahiz eta ebaluazioa burutzeko erabiliko den $Berria_{Test}$ dokumentu-sorta garapenean erabilitakoaren ezberdina izan, biak iturri berekoak dira. Horrek sistemari datu horien gaineko abantaila eman liezaioke.

Ebaluazio ahalik eta errealena egiteko asmoz, $Gara_{test}$ bigarren dokumentu-sortaren gainean ere ebaluatu dugu sistema. Bigarren datu-sorta hori ere kazetaritza-arlokoa da, eta, alde horretatik, onartu beharra dago arlo arteko ebaluaziorik egiteko aukerarik ez duela ematen.

Aitzik, iturri ezberdinetatik datozenez, abagune paregabea eskaintzen digu sistemaren errendimendua modu sendoagoan haren garapen-eremutik kanpo ebaluatu ahal izateko.

Ildo horretatik, ebaluazio hori bereziki kontuan hartzekoa da corpusetan oinarritutako lexikoen kasuan. Izan ere, atalean azaldu bezala, itzulitako L_{Trans} lexikoa, corpusarekin erlaziorik ez duen heinean, hobeto egoki liteke dokumentu berriak sailkatzerakoan.

Edonola ere, ezin dugu ahaztu, lexikoa edozein dela ere, sailkapena markatzen duen α atalasearen optimizazioa $Berria_{Train}$ dokumentu-sortaren gainean burutu dela, eta horrek ere azken emaitzan eragina izan dezakeela.

4.5 POS informazioa

Literaturan aipatzen da izen, aditz, adjektibo eta adberbioak direla iritzia adierazten duten hitzak. (Yu eta Hatzivassiloglou, 2003) lanak aditz, adjektiboak eta adberbioak hartu zituen hitzik adierazgarrienak bezala. Subjektibotasuna adierazten duten izenen bila jo dute bereziki batzuek (Riloff et al., 2003), eta beste zenbaitek ere (Das eta Bandyopadhyay, 2009b), erabili izan dute POS etiketen informazioa subjektibotasunaren sailkapenerako.

Iritzien sailkapenarekin lotutako (Kouloumpis et al., 2011) lanak, berriz, zalantzan jartzen du kategoria gramatikalek egiten duten ekarpena alor horretan. Gure asmoa da lexiko subjektiboak sortzean informazio hori kontuan hartzeak duen eragina aztertzea.

Informazio hori gure sisteman integratzeko, hitzak bere kategoriarekin lotuta egongo dira, subjektibotasun-lexikoak eraikitzeko garaian. Ondorioz, lehen lema bakarrean bildutako hitz guztiak orain elementu ezberdinetan banatu daitezke.

Adibidez, "izan" lema zuten hitz guztien maiztasunak sarrera bakarrean biltzen ziren lexikoan. Orain, aldiz, "izan_ADI" eta "izan_ADL" bi sarreratan banatuta daude. Horrela, lexikoak hitzen errepresentazio errealagoa izan dezan lortu nahi dugu.

4.5 taulak erakusten du "izan" lemak zer nolako pisuan duen kategoriarik gabeko lexikoetan eta kategoriadun lexikoetan.

L_{LLR}		$L_{LLR-KAT}$	
izan	1963.956	izan_ADT	3513.070
		izan_ERL	2260.544
		izan_IZE	63.715
		izan_ADL	59.957

Taula 10: "izan" lemaren pisuak LLR neurriarekin lortutako lexikoetan, POS informazioa kontuan hartuta eta kontuan hartu gabe

Kategoria gramatikalaren eragina aztertzeko, esperimentuak $Berria_{Test}$ ebaluazio-corpusaren gainean egin dira, eta corpusetan oinarritutako lexikoak erabili dira esperimentu horietan. Horrez gain, Yu eta Hatzivassiloglou (2003) autoreek proposatutako konbinazioa aztertu dugu, eta kategoria ezberdinek egiten duten ekarpena ere kuantifikatu.

5 Emaitzak

Datozen lerroetan, aurreko atalean deskribatutako esperimentuen emaitzak aurkeztuko ditugu.

Ebaluazioan zehar, gure sailkatzailearen errendimendua neurtzeko bi metrika erabili ditugu, literaturan oso erabiliak izan direnak: zehaztasuna (accuracy) eta F-score delakoa.

Zehaztasuna: sailkatutako elementu guztien artean (populazio osoa), zuzentasunez zenbat sailkatu diren adierazten du. Asmatze-tasa ere esaten zaio, eta honako formula hau du:

$$\text{Zehaztasuna} = \frac{\text{subjektibo zuzenak} + \text{objektibo zuzenak}}{\text{subjektibo zuzenak} + \text{subjektibo okerrak} + \text{objektibo okerrak} + \text{objektibo zuzenak}} \quad (3)$$

F-score: doitasunaren (precision) eta estalduraren (recall) arteko batazbesteko haztatua da neurri hau. Doitasunak adierazten du sistemak kategoria batean sailkatu dituenen artean zenbat sailkatu diren zuzentasunez. Estaldurak, berriz, neurtzen du kategoria batean hasieran zeudenetatik sistemak zenbat sailkatu dituen zuzen. Ondorengo ekuazioetan, hiru neurriak islatzen dira, baina gure kasuarekin lotuta:

$$Fscore = 2 \cdot \frac{\text{doitasuna} \cdot \text{estaldura}}{\text{doitasuna} + \text{estaldura}}$$

non

$$\text{Doitasuna} = \frac{\text{subjektibo zuzenak (objektiboak)}}{\text{subjektibo zuzenak (objektiboak)} + \text{objektibo okerrak (subjektiboak)}}$$

eta

$$\text{Estaldura} = \frac{\text{subjektibo zuzenak (objektiboak)}}{\text{subjektibo zuzenak (objektiboak)} + \text{subjektibo okerrak (objektiboak)}}$$

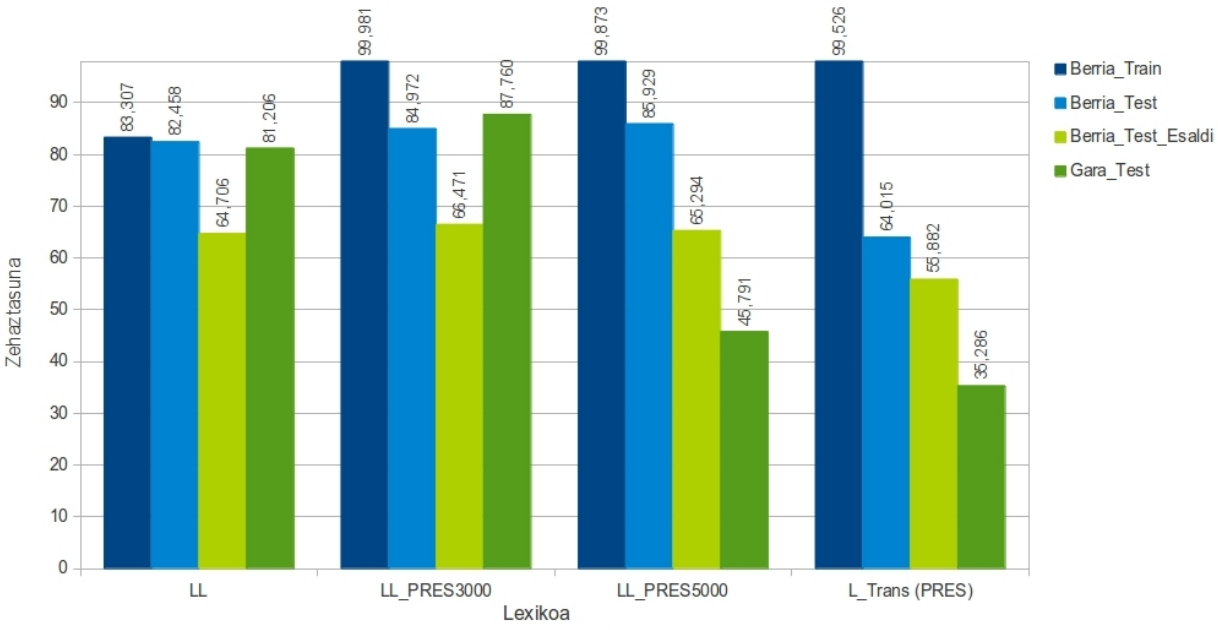
Hurrengo alderdi hauek ebaluatu ditugu:

- subjektibotasun-lexikoak sortzeko metodoen egokitasuna: lexikoaren itzulpena edo automatikoki eraikitako lexikoak ebaluatu ditugu, eta aztertu dugu testu-bilduma ezberdinetan zeinek maximizatzen duen sistemaren zehaztasuna.
- Sistemak testu laburragoen sailkapen-ataza baten aurrean duen errendimendua: zehaztasuna eta fscore neurrien bidez, ebaluatu dugu sistemak zer nolako portaera duen iturri ezberdinetako bildumetan.
- Sistema orokorraren errendimendua dokumentu-bilduma ezberdinen aurrean: hemen ere, zehaztasuna eta fscore neurrien bidez, ebaluatu dugu sistemak zer nolako portaera duen iturri ezberdinetako bildumetan.
- Hitzen informazio morfosintaktikoaren eragina: kategoria gramatikalaren informazioa baliatzeak emaitzetan zer eragin duen aztertu dugu, eta zehaztasun-balioetan adierazi.

5.1 Subjektibotasun-lexikoak

Subjektibotasun-lexikoak sortzeko metodologiak ebaluatzeko, corpusen arteko dibergentzian oinarritutako lexikoak proiekzio-lexikoaren ezaugarrietara egokitu ditugu. Aurrez 4.2.3 atalean azaldu bezala, estrategia izan da bigarren metodologiako lexikoen subjektibotasun-balioak kontuan ez hartzea. Horrela, lexiko guztiak subjektibititatea detektatzeko algoritmoa modu berean erabiltzera behartu ditugu.

Bestetik, ebaluazioa gauzatzeko, presentzian oinarritutako $sub(t)$ funtzioaren aldaera erabili da, eta prestatutako datu-sorta guztiak sailkatu dira. Ebaluazioan, presentzia darabilten hiru lexikoak sartu ditugu: $L_{LL_PRES3000}$ eta $L_{LL_PRES5000}$ corpusetan oinarritutakoak eta L_{Trans} proiekzioa. 1 irudian zehaztasun-datuak ematen dira. Halaber, irudiak L_{LL} lexikoarekin (subjektibotasun-pisuak kontutan hartzen ditu) lortutako emaitzak ere erakusten ditu erreferentzi gisa.



Irudia 1: Itzulitako L_{Trans} eta automatikoki sortutako lexikoekin lortutako zehaztasun datuak.

L_{Trans} ingelesetik proiektatutako lexikoa $L_{LL\_PRES5000}$ hutsetik sortutakoak baina okerragoa da kasu guztietan. $Berria_{Test}$ bilduman batez ere, % 21 aldea dago zehaztasunean, eta $Gara_{Test}$ nahiz esaldi-bilduman ere, aldea nabarmena da. Pentsa liteke alde hori lexikoaren estaldura handiagoak eragiten duela, baina $L_{LL_PRES300}$ lexikoak $L_{LL_PRES500}$ lexikoaren emaitza oso antzekoak lortzen ditu, Berria egunkaritik eratorritako bildumetan.

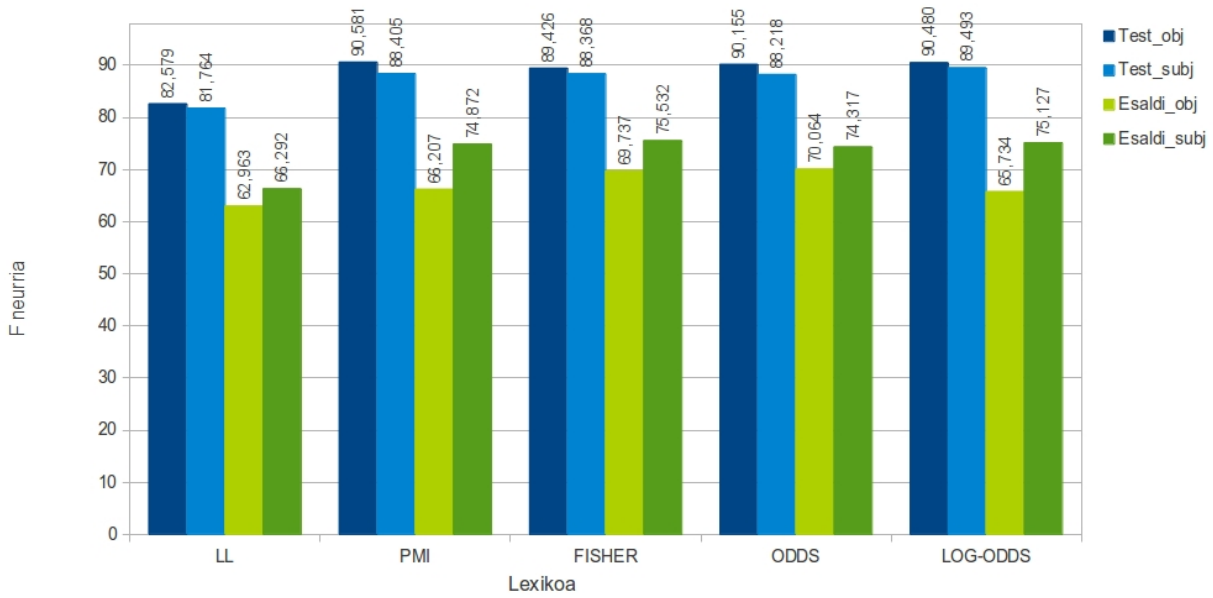
$Gara_{Test}$ bildumari dagokionez, aipatu behar da, corpusetik erauzitako $L_{LL_PRES500}$ lexikoak zein L_{Trans} proiekzioak oso zehaztasun baxuak lortzen dituztela. Dokumentuen

sailkapena aztertu dugu horretarako arrazoi baten bila, eta ikusi dugu dokumentu objektiboen doitasuna % 0 dela, hau da, dokumentu bakar bat ere ez da ongi sailkatzen. Dokumentu subjektiboekin erdietsitako zehaztasuna, ordea, nahiko altua da bi kasuetan (%91,6 $L_{LL_PRES500}$-ren kasuan eta % 70,6 L_{Trans}-renean). Beraz, bistakoa da $Berria_{Train}$ bildumarekin zehaztutako α atalasea zorrotzegia dela kasu horietan.

Azkenik, aipagatzekoa ere bada $L_{LL\_PRES3000}$ lexikoarekin lortutako emaitza. $Berria_{Test}$ eta *Berria_esaldi* datu-sorten gainean, 5.000 sarrerako lexikoaren pareko emaitzak lortzen ditu; baina $Gara_{Test}$ dokumentu bildumarekin, % 87ko zehaztasuna lortzen da. Kasu horretan, atalasearen zorroztasunaren eraginik ez dago. Lexikoak berak $Berria_{Test}$ bildumako dokumentuekin lortutakoa ere gainditzen du, eta pisuak kontutan hartzen dituen lexikoa ere gainditzen du. Aztertu beharko genuke subjektibotasun-pisuak dituzten lexikoetan sarrera-kopurua mugatzeak nola eragiten duen.

5.2 Dokumentu-mailako vs. esaldi-mailako subjektibotasuna

2 irudiak fscore balioak erakusten ditu corpusetan oinarritutako metodoan sortutako lexikoentzat. Orokorrean, fscorearen galera bat dago lexiko guztien kasuan, esaldien sailkapena burutzean dokumentuen sailkapenarekiko. Emaitza hori esperotako da, kontutan izanda sistema testu luzeekin optimizatu dugula.



Irudia 2: Berria egunkariko Dokumentu- eta esaldi-sorten gainean lortzen diren fscore datuak Automatikoki lortutako lexikoetarako

Lexikoei dagokienez, L_{LL} lexikoak ditu emaitzarik okerrenak. % 62 inguruko f-score balioa lortzen du dokumentu objektiboen sailkapenean, eta % 66 ingurukoa, aldiz, doku-

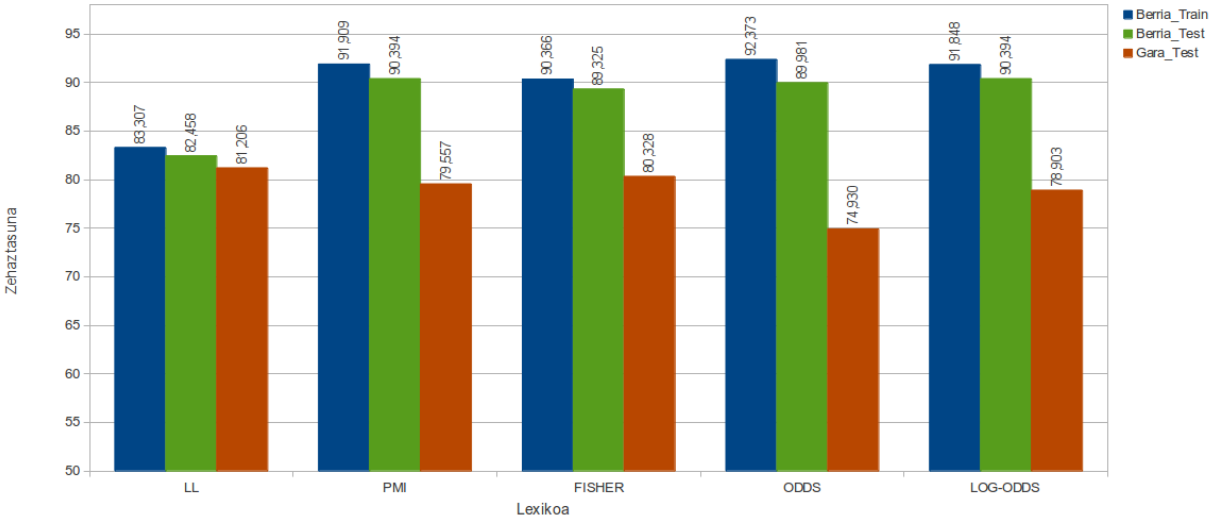
mentu subjektiboen sailkapenean, eta emaitza horiek L_{odds} lexikoak dituen % 70 eta % 74 balioetatik nahiko urrun daude.

Lexiko guztiek jasandako galerak dokumentuen sailkapenean lortutako emaitzei dagozkienez, oso antzekoak dira. Horrenbestez, ezin dugu esan lexiko bat besteak baino hobe egokitu denik ataza honetara.

Azkenik, aipatzekoa da ere esaldi objektiboen sailkapena bereziki okertzen dela, eta dokumentu-mailan datu objektiboen sailkapena hobea bazen ere, esaldi-mailan datu subjektiboak dira emaitzarik onenak lortzen dituztenak.

5.3 Berria vs. Gara

Dokumentu-mailako sailkatzailearen errendimendua bi dokumentu-sortaren gainean ebaluatu da: $Berria_{Test}$ eta $Gara_{Test}$. Lehehengoak 5.330 dokumentu ditu, eta bigarrenak, aldiz, 9.338 dokumentu, eta kasu bietan % 50 subjektiboak dira, eta beste % 50, objektiboak.



Irudia 3: Berria eta Gara dokumentu-sorten gainean lortzen diren zehaztasun datuak.

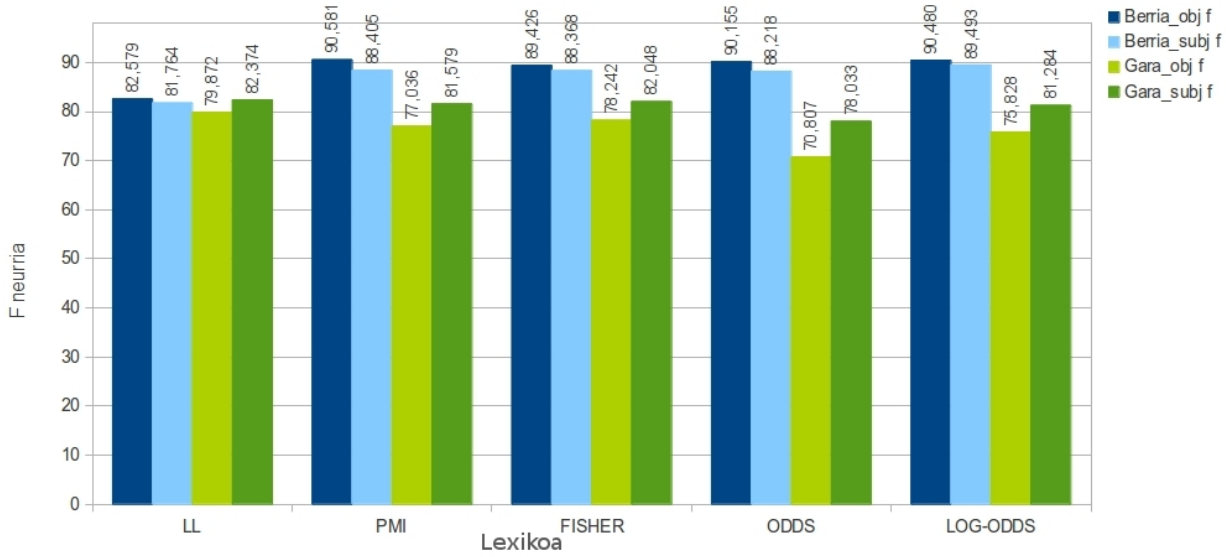
3 irudian sortutako lexiko guztiekin lortutako zehaztasun-datuak ematen dira, ebaluazioko bi bildumekin lortutakoak zein garapen fasean erabilitako $Berria_{Train}$ bilduman izandakoak. $Berria_{Test}$ eta $Gara_{Test}$ bildumen artean ageri diren aldeak aztertzerakoan, aipatu behar da lexiko gehienek galera nabarmena dutela $Gara_{Test}$ bilduma sailkatzerakoan.

Laburbilduz, garapen-bildumarekin nahiz $Berria_{Test}$ bildumarekin emaitzarik onenak % 90 inguruko zehaztasuna lortu duten lexikoak dira, % 10 edo gehiagoko galera izan dutelarik. Horien artean, L_{Fisher} lexikoa egokitzen da hobekien, eta horrek ere % 9ko

galera du. $L_{Log-odds}$ lexikoa, ordea, Berriako bilduman onena izatetik ia zehaztasunik okerrena izatera pasatzen da.

Lexiko guztien artean, gehien nabarmentzen dena L_{LLR} da. Izan ere, $Berria_{Test}$ bildumarekin emaitza okerrena eman duen neurria izan arren, zehaztasun ia berdina lortzen du bilduma batetik bestera igarotzean, $Gara_{Test}$ bilduman lortzen duelarik emaitzarik onena.

Beraz, emaitza horiek ikusita, esan dezakegu LLR elkartze-neurriarekin sortutako garapen-corpusarekiko nahiko independentea dela, eta dokumentu-sorta berrietara ongi egokitzen dela. Lan honen mugetatik at geratzen da beste arloetara ere ongi egokitzen den aztertzea.



Irudia 4: $Berria_{Test}$ eta $Gara_{Test}$ dokumentu-sorten gainean lortzen diren f neurriak bi kategorientzako.

Kategoria bakoitzerako lortutako fscore neurriak aztertzen baditugu (ikus 4 irudia), ikus dezakegu $Berria_{Test}$ bilduman bi kategoriatako dokumentuak nahiko fscore parekoarekin sailkatzen direla.

$Gara_{Test}$ bilduman, aldiz, bereziki dokumentu objektiboen sailkapenean egiten du behera fscore-k. Jaitsiera bortitz hori pairatzen ez duen lexiko bakarra L_{LLR} da.

5.4 POS informazioa

Kategoria gramatikal ezberdinek subjektibotasunaren detekzioan duten eragina balioesteko, hainbat konbinazio aztertu dira. Baseline gisa, POS informaziorik ez duen sistema erabili dugu, eta ebaluazioa $Berria_{Test}$ dokumentu-sortaren gainean burutu dugu, corpusetan oinarritutako lexiko guztiak erabiliz.

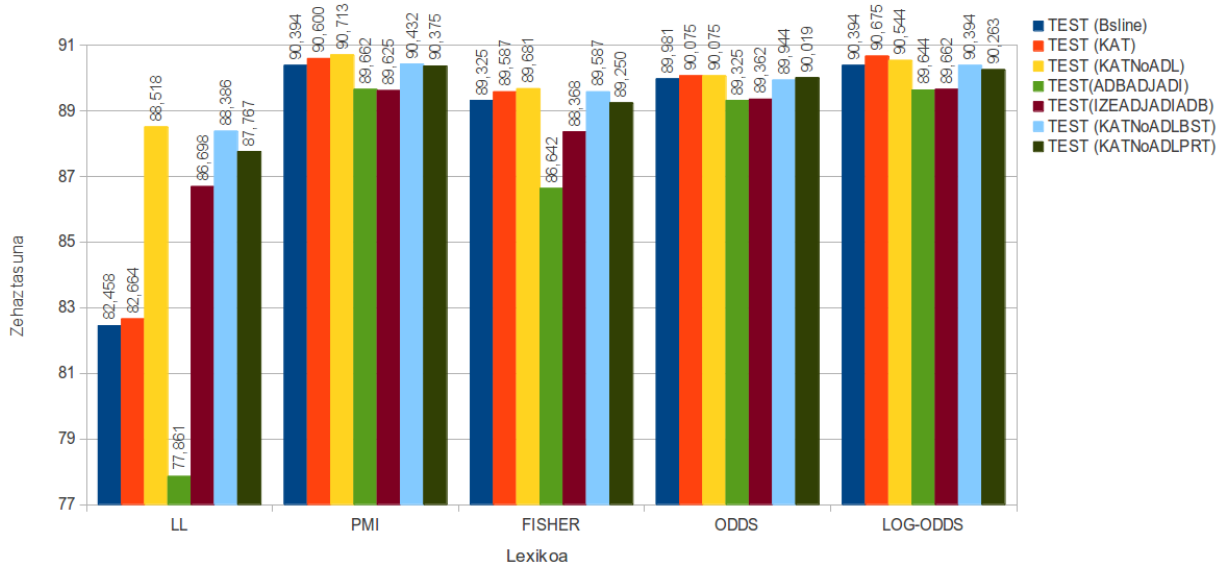
HAP masterra

5 irudiak konbinazio esanguratsuenekin $Berria_{Test}$ dokumentu-sortaren gainean lortutako zehaztasun berak ematen ditu. Orokorrean, kategoria-informazioarekin lortutako emaitzen arteko aldea oso txikia da, eta ondorio argiak ateratzea zaila da. Hala ere, LLR neurriaz bestelako neurriekin lortutako emaitzek joera berdina erakusten dute. Dirudienez, neurri hori POS informazioarekiko sentikorragoa da, besteak baino. LLR neurriaren emaitzei dagokienez, aparte komentatuko ditugu.

Oro har, zertxobait laguntzen duela erakusten dute emaitzek, hitz guztien POS informazioa gehitzeak hobekuntza txiki bat baitakar LLR lexikoaz bestelako lexiko guztiekin. Kategoria zehatzei erreparatuz gero, kategoria batzuetako hitzak baztertzen hasten garenean emaitzak beherantz doaz kasu guztietan.

Kategoria bakoitzaren portaera aztertuta, ikusten da izenek (IZE), adjektiboek (ADJ) eta adberbioek (ADB) egiten dutela ekarpenik handiena, beste ikerketa-lan batzuek ondorioztatutakoarekin bat etorriz. Aditzek (ADI) ere egiten dute beren ekarpena, baina aurreko hiru kategoriena baino txikiagoa da. Era berean, aditz laguntzaileen (ADL) ekarpena oso baxua da, eta horiek kentzeak oso galera txikia dakar kasu guztietan.

Amaitzeko, esan behar PRT edo BST moduko kategoriek -corpusean oso dentsitate baxua dutenak- ekarpen txiki bat behintzat egiten dutela. 5 irudian ikus daitekeen bezala, KATNoADLPRT eta KATNoADLBST zutabeek galera txiki daukate dute neurri guztietan, KATNoADL zutabearekin alderatuta. Izan ere, KATNoADL zutabeak kategoria horietako hitzak erabiltzen ditu, eta beste biek ez.



Irudia 5: Berria eta Gara datu-sorten gainean lortzen diren zehaztasun datuak.

LLR neurriak aparteko sentikortasuna erakusten du POS informazioarekiko. Beste kasuetan bezala, kategoria gramatikalaren informazioa besterik gabe erabiltzeak hobekuntza txiki bat besterik ez dakar baseline sistemarekiko.

HAP masterra

Kategoriak aukeratzerakoan, aldiz, hobekuntza oso nabarmena dauka, baselinearekiko ia 6 puntu irabazten baituta, eta zehaztasuna ere ia %89ra iristen da, beste neurrien emaitzetara asko hurbiltzen delarik. Bestalde, bereziki aipagarria da aditz laguntzaileek (ADL) duten eragin negatiboa. Izan ere, horiek kentzeak eragiten du hobekuntzarik handiena.

Aipagarria da ere, beste elkartze-neurriekin baino nabarmenagoa delako, (Yu eta Hatzivassiloglou, 2003) lanean emaitzarik onenak lortzen zituen aditzen, adjektiboen eta adberbioen konbinazioak (ADBADJADI zutabea 5 irudian) ez duela emaitza onik ematen gure kasuan; baina izenek, aldiz, oso ekarpen garrantzitsua egiten dute (IZEADJADBADI zutabea.).

Hitz gutxitan, emaitza horiek -labelsec:ema:TestDataEzberdinak atalean ikusitakoekin batera- LLR neurria ez dela baztertu behar ondorioztatzera garamatzate.

6 Ondorioak eta etorkizuneko lanak

Lan honetan, euskarazko testuetan subjektibotasuna detektatzeko teknikak landu dira, iritzia adierazten duten dokumentuak eta esaldiak identifikatzeko helburuarekin. Horretarako, gainbegiratu gabeko sistema bat garatu dugu, subjektibotasun-lexiko batean oinarritutako subjektibotasun-informazioa ustiatzen duena testu-zati bateko subjektibotasuna neurtzeko.

Subjektibotasuna identifikatzeko sistema automatiko bat garatzeko, ezinbestekoa dugu dagoeneko sailkatuta dauden testuen edo esaldien erreferentziazko corpus bat sortzea. Nahiz eta printzipioz kalitate hobea espero daitekeen, corpus bat subjektibotasun-informazioarekin eskuz etiketatzearen kostua oso handia da, eta gure kasuan, ezin dezakegu gure gain hartu. Hori dela eta, lan honetan dokumentu-mailako subjektibotasuna duen kazetaritza-corpus bat automatikoki bildu dugu, eta corpus hori sistema garatzeko nahiz ebaluatzeko erabili da.

Horrez gain, esaldi-mailako subjektibotasunaren detekzioa ebaluatzeko, beste bi datu-sorta sortu dira: bata, erreferentziazko corpus horretatik abiatuta, eta bestea, beste iturri batetik bildua, kazetaritza-arlokoa ere bai. Bi datu-sorta horiek eskuz sailkatuta daude.

Dokumentu-mailako detekzioari dagokionez, oso emaitza onak izan ditugu, % 90 inguruko doitasunak lortuz. Esaldi-mailako detekzioari dagokionez, ordea, doitasuna jaitsi egin da dokumentu-mailako detekzioarekin alderatuz. Dena den, asmatze-tasa nahiko altuak lortzen dira, eta kontuan izan behar da dokumentu-mailan lan egiteko optimizatutako sistema erabili dugula, inolako aldaketarik egin gabe.

Halaber, proposatutako metodoa iturri ezberdinetako datuetara nola egokitzen den ere aztertu dugu. Emaitzak eskasagoak dira, % 10eko galera bat izan dugularik. Horrek esan nahi du automatikoki sortutako subjektibotasun-lexikoek horiek erauzteko erabili diren

corpusarekiko mendekotasun altua dutela. Ondorio hori are gehiago sendotzeko, ingelesetik itzulitako lexikoen kasuan ere, ikusten dugu galera txikiagoa dela. Aurrera begira, mendekotasun hori murriztuko duten bideak aztertu beharko lirateke; besteak beste, itzulitako lexikoak eta corpusetatik erauzitakoak konbinatzea.

Subjektibotasun-lexikoari dagokionez, lexikoa sortzeko bide ezberdinak aztertu dira, eta sortutako lexikoen egokitasuna ebaluatu da. Egindako esperimentuen arabera, beste hizkuntza bateko lexiko bat hartzea baino egokiagoa da euskararako lexiko berriak hutsetik sortzea. Izan ere, sortutako lexikoek estaldura handiagoa dute, eta hobeto egokitzen dira sailkapen-ataza ezberdinetara.

Hala ere, itzulpen-prozesuan hobekuntzak egin daitezke, hala nola adieren desanbiguazioa hobetu eta itzulitako sarreren estaldura handitu. Esan beharrik ez kalitate handiagoko baliabide batek lan honek lortutako emaitzak alda litzakeela.

Lexikoa hutsetik sortzeko, metodo erabat automatikoaz baliatu gara, hainbat neurri ezberdin erabiliz. Horrez aparte, hainbat elkartze-neurriren araberako lexikoak sortu dira, emaitza ezberdinak lortuz. Emaitzarik onenak lortzen dituen neurria log-odds da orokorrean.

Bestalde, aipagarria da zeharo zabalduta dagoen LLR dela aztertutako neurrien artean emaitzarik okerrenak lortzen dituena, garapenean nahiz $Berria_{Test}$ ebaluazio bildumarekin. Itzulitako lexikoa gainditzen duen arren, % 8 galera du, gainerako lexiko automatikoekin lortutako emaitzekin alderatuz. Haatik, $Gara_{Test}$ bildumarekin ebaluatzerakoan, LLR-k oso portaera sendoa erakusten du, apenas galerarik gabe, eta zehaztasunik altuena lortzen du. LLR-rekin sortutako lexikoak, gainera, beste guztiek baino hobekuntza handiagoa lortzen du kategoria gramatikalaren informazioa gehitzen denean.

Ondorioz, sakonago aztertu beharko litzateke neurri horiek nola egokitzen diren testuetako ezaugarri linguistiko gehiago erabiliz gero. Lortutako emaitzen arabera, LLR elkartze-neurria litzateke egokiena lexikoak sortzeko, dokumentu-bilduma berrietara hobekien egokitzen dena delako.

Etorkizunera begira, lexikoa eskuz sortzeak kostu izugarria izango luke; baina interesgarria litzateke automatikoki sortutako lexikoak eskuz zuzentzea, eta kostu onargarri baten truke hobekuntza lortu ote daitekeen aztertu, Saralegi eta San Vicente (2012) egileek polaritate lexikoekin egindako lanaren bidetik.

Hitzen kategoria gramatikalak subjektibotasunaren detekzioan duen eraginari dagokionez, zalantzazkoa da lortutako emaitzetatik ondorio garbirik atera daitekeen. Elkartze-neurri gehienekin lortutako emaitzetan hobekuntza txiki bat lortu bada ere, hobekuntza hori % 0,35ekoa da kasurik onenean. LLR elkartze-neurriak bakarrik erakusten du hobekuntza nabarmen bat (% 6), POS informazioari esker. Kasu horretan, benetan esan liteke informazio hori erabiltzeak iritzia adierazten duten testuak identifikatzen laguntzen duela.

Kategoria batzuek besteek baino laguntza handiagoa ematen duten arren, kategoria gehienetako hitzak erabiltzea da egokiena. Gure esperimentuek erakutsitakoaren arabera,

aditz laguntzaileek zarata sortzen dute, eta LLR elkartze-neurriaren kasuan, bereziki nabarmena da. Horiez aparte, ez dago "zarata" sartzen duen kategoriarik, denek egiten dute ekarpen bat subjektibotasunaren detekzioan. Orobat, aipatzekoa da izenek egiten duten ekarpena. Elkartze-neurri guztiekin orokorrean, eta LLR elkartze-neurriarekin bereziki, zehaztasuna galtzen da izenak erabiltzen ez direnean.

Azkenik, aipatu beharra dago lan hau polaritatea detektatzeko ikerketa-marko baten barruan burutu dela. Subjektibotasuna detektatzeak polaritatearen detekzioa burutzen lagunduko digu, eta lan hau bide horretan emandako lehen pausoa da. Hemen egindako esperimentuetan sakondu beharra dago, hortaz. Dokumentu-mailako sailkapenean emaitza nahiko onak lortu dira, baina badago hobetzeko tartea. Era berean, esaldi-mailako sailkapenari ere ezinbestean heldu behar diogu, Twitter edo Tumblr[7] moduko microblogging zerbitzuetatik iritziak erauziko baditugu, edo webguneetako iruzkin-jarioetatik, esaterako.

Etorkizun hurbileko lan-ildoetan sartzen da ikasketa automatikoan oinarritutako sailkatzaile gainbegiratu bat garatzea. Horrez gain, ezin dugu ahaztu egungo sistema kazetaritzaren arlotik kanpo nola moldatzen den ere aztertzeke dagoela.

Aipatutako asmo horiek guztiak gauzatzeko, ezinbestean entrenamendurako nahiz ebaluaziorako datu-sortak beharko ditugu. Zentzu horretan, blogetatik eta Twitterretik ateratako testuekin eta liburuen eta pelikulen inguruko kritikekin osatutako esaldi-sortak etiketatzeko asmoa dugu.

Amaitzeko, subjektibotasunaren detekzioan lagun dezaketen beste ezaugarri batzuk ere aztertzeko leudeke, egungo sailkatzailean nahiz beste batean integratzeko; besteak beste, hitzen ordena, hitz subjektiboen habiatze-maila sintaktikoa edo diskurtsoaren analisia.

[7] https://www.tumblr.com/

Erreferentziak

Itziar Aduriz eta Arantza Díaz de Ilarraza. Morphosyntactic disambiguation ands shallow parsing in computational processing of basque. 2003.

Cecilia Ovesdotter Alm, Dan Roth, eta Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 579–586, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Krisztian Balog, Gilad Mishne, eta Maarten de Rijke. Why are they excited?: identifying and explaining spikes in blog mood levels. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, page 207–210, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

C. Banea, R. Mihalcea, eta J. Wiebe. Multilingual sentiment and subjectivity analysis. *Multilingual Natural Language Processing*, 2011.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, eta Samer Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, page 127–135, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

Johan Bollen, Huina Mao, eta Xiao-Jun Zeng. Twitter mood predicts the stock market. *1010.3003*, October 2010.

Rebecca F Bruce eta Janyce M Wiebe. Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5:187–205, June 1999. ISSN 1351-3249.

P. Chaovalit eta L. Zhou. 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*, page 112c, 2005. ISBN 0769522688.

Beatrice Daille. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Technical report, 1995.

A. Das eta S. Bandyopadhyay. Subjectivity detection in english and bengali: A CRF-based approach. *Proceeding of ICON*, 2009a.

A. Das eta S. Bandyopadhyay. Theme detection an exploration of opinion subjectivity. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009*, pages 1 –6, September 2009b. doi: 10.1109/ACII.2009.5349599.

HAP masterra

Andrea Esuli eta Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, page 617–624, New York, NY, USA, 2005. ACM. ISBN 1-59593-140-6.

Andrea Esuli eta Fabrizio Sebastiani. SENTIWORDNET: a publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422, 2006.

S. Evert. The statistics of word cooccurrences. *Word Pairs and Collocations. Phil. Diss. Institut f" ur maschinelle Sprachverarbeitung. Stuttgart*, 2005.

Vasileios Hatzivassiloglou eta Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

Minqing Hu eta Bing Liu. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*, pages 755–760, San Jose, California, 2004. AAAI Press / The MIT Press. ISBN 0-262-51183-5.

Nobuhiro Kaji eta Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, EMNLP-CoNLL'07, pages 1075–1083, Prague, Czech Republic, 2007. Association for Computational Linguistics.

Hiroshi Kanayama eta Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 355–363, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6.

J. Kessler, M. Eckert, L. Clark, eta N.Ñicolov. The ICWSM 2010 JDPA sentiment corpus for the automotive domain. In *International AAAI Conference on Weblogs and Social Media Data Challenge Workshop*, 2010.

A. Kilgarriff. Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133, 2001.

Soo-Min Kim eta Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

E. Kouloumpis, T. Wilson, eta J. Moore. Twitter sentiment analysis: The good the bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

HAP masterra

Isa Maks eta Piek Vossen. Building a fine-grained subjectivity lexicon from a web corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, eta Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

R. Mihalcea, C. Banea, eta J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, volume 45, page 976, 2007.

Bo Pang eta Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Bo Pang, Lillian Lee, eta Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Veronica Perez-Rosas, Carmen Banea, eta Rada Mihalcea. Learning sentiment lexicons in spanish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, eta Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

R. Quirk, S. Greenbaum, G. Leech, eta J. Svartvik. *A comprehensive grammar of the English language*. Pearson Education India, 1985.

Paul Rayson eta Roger Garside. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora - Volume 9*, WCC '00, page 1–6, Hong Kong, China, 2000. Association for Computational Linguistics. ACM ID: 1117730.

Ellen Riloff eta Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112, 2003.

Ellen Riloff, Janyce Wiebe, eta Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, pages 25–32, Edmonton, Canada, 2003.

X. Saralegi eta M. López de Lacalle. Comparing different approaches to treat translation ambiguity in CLIR: structured queries vs. target co-occurrence based selection. In *DEXA*

Workshops, Proceedings of the 6th international workshop on Text-based Information Retrieval, pages 398–404, 2009.

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

Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 417, Philadelphia, Pennsylvania, 2002.

Xin Wang eta Guo-Hong Fu. Chinese subjectivity detection using a sentiment density-based naive bayesian classifier. In *2010 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 6, pages 3299 –3304, July 2010. doi: 10.1109/ICMLC.2010.5580700.

J. Wiebe, T. Wilson, eta M. Bell. Identifying collocations for recognizing opinions. In *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, page 24–31, 2001.

Janyce Wiebe eta Rada Mihalcea. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, page 1065–1072, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, eta Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210, 2005. ISSN 1574-020X.

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

Theresa Wilson, Janyce Wiebe, eta Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005b. Association for Computational Linguistics.

Theresa Ann Wilson. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States.* ProQuest, 2008. ISBN 9780549733249.

Hong Yu eta Vasileios Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

N. Yu eta S. Kübler. Filling the gap: Semi-supervised learning for opinion detection across domains. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, page 200–209, 2011.