eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Agur-bertsoetako egitura diskurtsiboaren xerka.

Ikasketa automatikoaren zein erregela linguistikoen bidezko hurbilketa sailkatzailea

Mikel Osinalde Agirre

Tutoreak: Bertol Arrieta eta Mikel Lersundi

Laguntzaile eta eragilea: Aitzol Astigarraga

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua

lortzeko bukaerako proiektua

2013ko iraila

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

LABURPENA

Bat-bateko bertsogintzan agur-bertsoek atal beregaina osatzen dute, baina elkarrekin partekatzen duten ezaugarririk ba ote dute? Ba al dago ezaugarri horietan oinarrituta bertso mota horren egitura narratibo jakina edo askotarikoari antzematerik? Galdera horien erantzunaren bila dihardu honako lanak, zeinak azterketa eta sailkapen helburuak besarkatzen dituen. Horretarako, bide-buruan jarri eta bi norantzatan ekin diogu azterketa lanari. Batik bat testu-sailkapen automatikorako teknikekin osatu da esperimentuen multzo nagusia. Bagenuen aurreuste bat, hots, ezaugarri linguistikoak baliatuta emaitza onak lor genitzakeela. Alabaina, ikasketa automatikorako teknikek sailkatzaile egokiak bilatzeko eskain ziezaguketen laguntza egiaztatzen saiatu gara. Bertsoak egoki sailkatze aldera gai-kategoriatzat proposatu ditugu zenbait erreferente eta ideia; ondoren, aztergai ditugun hainbat bertso sorta analizatzeko baliatu ditugu aurrez zehaztutako kategoria horiek, egin ere, ikasketa automatikoko teknikak erabilita. Naive Bayes, k-NN, BayesNet, Support Vector Machines eta Decision Tree Learner sailkatze-algoritmoak hautatu ditugu eginkizuna burutzeko. Orobat, dimentsio-murrizte teknikak ezarri izan dira terminoek osatzen zuten eremua mehaztu eta egokitze aldera. Esperimentuetan erdietsitako emaitzek proposatutako hurbilketaren egokitasuna erabat baztertzen ez duten arren, bidea findu beharra iradokitzen dute eta aukera berrietarako abiapuntua zedarritzen. Bestalde, hasierako usteei oinarri sendoagoak ezartzeko aukera ere izan dugu. Etiketatutako bertso-puntuen behaketaren ostean, klaseak harrapatzeko zenbait hizkuntza-ezaugarri partekatu erauzi ditugu. Ondoren haien gainean erregelak sortu, eta azkenik, aurretiaz burututako lanaren pareko emaitzak lortzeko gai izan garela egiaztatu dugu.

Gako hitzak: Dokumentuen kategorizazioa, ikasketa automatikoa, diskurtsoaren egitura, dimentsioen murrizketa, erregela linguistikoak

ABSTRACT

Greeting-verses form an independent section in Basque improvised poetry, but do they share features among them? Is it possible, based on these characteristics, to find out specific, or varied, discourse patterns? The presented work tries to search an answer to these questions. The objective includes both, analysis and feature extraction of greeting verses, and classification based on those features. The main set of experiments had been composed using automatic text classification techniques. We assumed that taking into account linguistic features we could achieve good results. However, we tried to confirm whether machine learning techniques could help us finding the best classifiers. In order to classify the verses correctly, we propose some references and ideas as topic-categories; then, we used those pre-defined categories to analyze several verses. Machine learning techniques have been used to this purpose. Classification methods such as Naive Bayes, k-NN, BayesNet, Support Vector Machines and Decision Tree Learner have been selected. Dimensionality reduction techniques have been applied in order to reduce the term space. The results obtained in experiments do not exclude completely the suitability of the proposed approach; however, they suggest the need to redefine the task and to consider other approaches. On the other hand, we also had the opportunity to establish a solid basis for the initial opinion. After analyzing the phrases in the labeled verses, we extracted some linguistic features they share. Afterwards, we built some rules based on them, and finally, we have verified that we are able to achieve similar results with both machine learning and linguistic rules.

Keywords: Document categorization, machine learning, discourse pattern, dimensional reduction, linguistic rules

Aurkibidea

1. Sarrera

1.1. Lanaren nondik norakoak

Bat-bateko bertso ekoizpenean hauek antolatzeko eskuarki baliatzen den egiturarik ote den antzematen saiatu gara ondoko lanean. Ustez, nahiko ataza zedarritua burutzen den agur-bertsoen alorrera mugatu dugu ahalegina.

Agur-bertsoak aztergai hartuta bi urratsetan bereizi dugu egitekoa: a) Alorreko adituen laguntzaz agurren azterketa ahal bezain zabala egin dugu. Agur-bertsoetan ohiko ezaugarriak erauzi, bildu eta bateratu ditugu. b) Bigarren pausua, ezaugarri edo kategoria horien arabera testu-txatalak kategorizatzea izan da. Osatutako ikerketaren helburua ere bikoitza da: batetik, euskaraz osatutako bat-bateko agur-bertsoen balizko egitura narratiboa erauzi nahi genuke; bestetik, berriz, balizko egitura hori ikasketa-algoritmoen bidez zein adituen irizpenei jarraiki osatutako erregela linguistikoak baliatuta eskura ote litekeen ere egiaztatu nahi genuke, orobat, bi metodologien irismena balioetsita. Bi prozedura horiez diharduguneanan *Machine Learning (ML)* edo ikasketa automatikoaz eta erregela linguistikoek bideratutakoaz ari gara.

Honela antolatu dugu egindako lanaren berri ematen duen memoria: hastapenean, bertsolaritzari buruzko sarrera eskaintzen dugu, alegia, gure herrian bat-bateko jardunean lantzen den jarduera poetikoaren berri xumea ematen dugu; egin ere, ondorengo atazak hobeto uler daitezen. Jarraian, gramatika konputazionalaren eta testu-sailkatze iker-arloaren ingurukoez aritu gara gaingiroki. Bertan, diskurtsoaren egitura, dokumentuen errepresentazioa, ezaugarri murrizketa eta algoritmo bidezko sailkapena bezalako kontzeptuez ere zenbait ohar ematen dira. Ondoren, esperimentua abian jartzeko urratsez eta hura antolatzeko eraz dihardugu zehatz; izan ere, hurrengo atalean egindako esperimentu-saioak deskribatu eta horietan lortutako emaitzak erakusten baititugu, segidan, haietan aurkitutako alderdi esanguratsuez aritzeko. Azkenik, zenbait ondorio aurkezten ditugu eta etorkizunean osa daitezkeen lanen inguruko gidalerroak eskaintzen.

2. Ikergaia zehazten

2.1. Ahozko jarduerak aztertzeko, bertsolaritza oinarri

Aukeratu dugun corpusa eta bere testuingurua azaltzen hasteko ahozko tradizioaren unibertsaltasuna eta berezitasunak aztertu izan dituen John Miles Foley adituak nazioarteko ahozko inprobisatzaileen 2003ko topaketetan emandako hitzalditik jaso ditugu jarraian datozen lerroak.

> **Ahozko poesiak hizkuntzak bezala funtzionatzen du, baina zerbait gehiago ere eskatzen du**. Erraz ahazten dugu ahozko poesia ez dela gauzaki bat ez testu bat, ezarritako arauen arabera aldaketak jasan ditzakeen hizkera bizi bat baizik. Desberdintasunik handiena, eguneroko hizkerarekin alderatuz, hauxe da: ahozko poesiaren erregistro espezializatuek egitura gehiago erabiltzen dituzte, eta era konplexuago batean kodetutako esanahi idiomatikoak dituzte. Esapide horiek dira *zerbait gehiago* hori. [Foley, 2004]

Aztergai duguna ez bide da testu bat iker-arlo honen ikuspegitik, nolanahi ere, eta ausarkeria dirudien arren, hala bailitzan baliatu ditugu alor horretako zenbait emaitza (agur-bertsoak), izan ere, erregistro espezializatua bada eta egitura anitz baliatzen baditu, kodetutako esanahi idiomatikoak ezaugarri nagusitzat baditu, horiek aztergarri eta ezagugarri izango direlakoan baikara. Hurbilketa lan honen xedea behintzat hori izan bide da, tradizioaren kariaz iltzatuta geratu diren ohiko prozedurak eta bertsoa taxutzeko erabiltzen diren modu estandarrak azaleratzen ahalegintzea.

> ***Performancea***[1] **da gertakaria; tradizioa, berriz, gertakari horren testuingurua.** *Performance* bakoitzean desberdin ulertu behar da hizlariak dioena, eta entzuleek performance bakoitzera egokitu behar dute beren ulermena. Tradizioa da gertakariaren testuinguru edo erreferentzia; *performance* bakoitzaren *hutsuneak* betetzen ditu bertan gertatzen ari dena entzuleen esperientzia zabalagoarekin osatuz. [Foley, 2004]

Bigarren aipu honetan erreferentziek, hots, igorleak eta hartzaileak (zeinahi

1 Foleyren artikuluan ingelesezko *performance* hitza bertso saioetan ematen den elkarrizketa ekintzari, agerraldiari edo emankizunari ari zaio erreferentzia egiten. Ez bedi nahas, lan honetan aurrerago, emaitzen azalpenean eta IAko tresnen lorpenez ari garenean darabilgun terminoarekin, zeina bitarteko konputazionalen errendimenduari egokitzen zaion.

komunikazio ekintzatan bezala) partekatzen duten munduaren ezagutzari lotuta egonik aipatu gabe ere ulertzen diren gako horiek, zenbaterainoko indarra duten adierazten da. Hasiera batean hipotesitzat joko ditugun erreferenteok sailkatzaile automatiko bidez egiaztatzeko saiakera egingo dugu.

Esandakoaren harira, ahozko tradizioan oinarritutako teknikok oinarri orokorra, unibertsala izanik, herriz herri berariazko adierazpideak dituzte. Gure herrian, alegia, euskararen herrian bertsolaritza da egun adierazmolderik arrakastatsuena eginbide horretan. Horrek bideratu gaitu azterketan esparru jakin hau hartzera. Hona, Eusko Jaurlaritzako azterketa soziolinguistikoen sailak 2007an egindako inkestan bertsolaritzaren ezagutzari eta garrantziari buruz euskal herritarrek eskaini zuten ikuspegia:

> 2003ko EAEn bertsozaletasuna ez dago oso zabaldua (% 17) oso edo nahikoa zaleak dira), baina gehiengoa osatzen dute noizbait bertso saio bat entzun duten herritarrek (% 66k). Bertsolaritza, batez ere, herri kultura bezala definitu dute herritarrek (% 49k); haien iritziz, garrantzia du euskal kulturaren barruan (% 82k), eta erakunde publikoek orain arte beste (% 44k) edo gehiago (% 34k) babestu eta bultzatu beharko lukete. EAEko biztanleen erdiek baino gehiagok (% 58k) zer edo zer entzun zuten 2005eko abenduan izandako Bertsolari Txapelketa Nagusiko finalaz eta laurden batek (% 27k), inolako erantzunik iradoki gabe, 2006ko otsailean gogoratzen zuen haren irabazlea Andoni Egaña izan zela. [Prospekzio soziologikoen kabinetea, 2007]

Aipatutako inkesta soziolinguistikoa EAEko biztanle guztien artean egin zen. Erraz aurresan daiteke ordea, euskal hiztunen komunitatean are ezagutza zabalagoa azalduko zela eta nahiz eta garrantziari dagokion datua zinez esanguratsua den (% 82). 700.000 lagun inguruk osatzen dute bere hiztun komunitatea, esan nahi baita, herrialdeko biztanleen % 25a inguruk. Beraz, euskal hiztunen komunitatea txikia da hizkuntza nagusietako masarekin parekatuz gero. Gainera, egoera gutxituan dago bere jatorrizko lurraldean.

Esparru zabala da ahozkotasunarena, unibertsala ere bai, esan dugunez. Bertsolaritza euskararen komunitatean hiztunek sortutako haren adierazpide bat bada, guk baliatutako datuak gizatalde horretan oihartzun handia duen norgehiagoka eremukoak dira. Bat-bateko bertsolaritza lehiaketa oraindik ere indar handia duten Euskal kulturaren adierazpenetarikotzat hartu behar da. Gaingiroki bada ere, atazotan

aintzat hartu beharreko irizpide zenbait azaltzeko izan bitez ondoko lerrook.

Neurriz eta errimaz
kantatzea hitza,
horra, zer kirol-mota
den bertsolaritza!

Xabier Amuriza

Unean bertan osatzen diren bertsoak oinarritzat dituzten ekitaldiak eta lehiaketak oso dira ohikoak euskal herrietan. Halako emankizunetan bertsolari batek edo gehiagok inprobisatutako lanak egiten dituzte, hain zuzen ere, gai-jartzaile batek eskainitako abiaburuak nahiz gaiak aintzat hartuta. Argibideok jaso ostean bertsolariak zenbait segundo hartzen ditu, eskuarki minutua baino gutxiago, bertso bat taxutzeko, jakina, aurresandako bertso-egitura bati jarraiki, zeinak errima antolaera jakin bat ere hartzen duen. Ehunka doinuren artean aukeratzen dira melodiak. Tradizio luzekoak batzuk, berriki asmatuak beste hainbat. Bat-bateko bertsoak egituratzerakoan zenbait eskakizun formal hartu behar dira kontuan. Errima eta metrika bereizi ez daitezkeen elementuak dira abestu beharreko bertso inprobisatuetan. Nolanahi ere, bertsoaren zinezko balioa ez datza eskakizun teknikoon betetze mailan. Aitzitik, dialektika, erretorika nahiz poetika arloei dagokienean izan dezakeen balioan oinarritzen da bertsoaren kalitatea edo egokitasuna. Beraz, bertsolariak ideia eta pentsakizun eskerga modu originalean adierazteko gai izan behar luke, egin ere, aipatu betekizun teknikoek eragindako murrizketei aurre eginda. Oreka horretan datza preseski bertsoaren magia. Bertsolaritzako emanaldirik gogorrena, partaideak zorrotzen hartzen dituena *Bertsolari Txapelketa Nagusia* da, zeina lau urtean behin antolatzen den. Halako ekitaldietan bertsolari multzo batek elkarrekin lehian dihardu lehiaketa irabazi eta hurrengo lau urteetan denen artean txapeldun izendatu gisa aritze aldera.

Txapelketa parte hartzaileek bete beharreko hainbat atazatan bereizten da eta aipatu egitekootan eskatutako lana askotarikoa izan ohi da. Haatik, bat-bateko agur-bertsoekin ekiten diote saioei beti, eta baita amaitu ere. Inork gai jakinera mugatu gabe, libre aritzen diren bertsokeran osatzen dituzte puntuak. Bertso modu hau izaten da

bertsolariari berak nahi duena zuzenean adierazteko modua eskaintzen dion bakarra. Aurrerantzean, lehiaketan zehar gai-jartzaileak jarriko ditu bertsolariak abiaburutzat edo jarraibidetzat hartu bide dituen gaiak. Gainera, bertsoen metrika eta ale kopurua ere zehaztu egingo zaizkio.

Goiko arrazoien kariaz, agur-bertsoak bereziki interesgarritzat jo ditugu bertsolariek beraien narrazioetan darabilten ustezko egituraren bila aritzeko eta hura aztertzeko.

2.2. Ikerketa ikuspegiak, zeruertza eta usteak

Agur-bertsoetan ohiko arrazoibide ildorik dagoen antzematea litzateke lan honen xedea, beraz. Bestela esanda, bertsolariek partekatzen duten diskurtsoa egituratzeko modu jakinen bila dihardugu. Helburu horretarako ezinbestekoa genuen esanahiarekin zerikusia duten hainbat aldagai aztertzea. Bagenuen esanahi linguistikoari soilik begiratzea, bertsoetako azterketa unitatetzat hitzak hartu, semantikaren ikuspegitik etiketatu eta bildutako informazio multzoaren ondorioak behatu. Haatik, esanahiaren azterketa hori testuinguruaren arabera egiteak aukera gehiago zemaigun bat-bateko jardunean gertatzen diren alderdi gehienak biltze aldera (erreferentziak, inferentziak, bertsolariak esaten duena, asmoa, norbere egoera, testuingurua). Pragmatikak gizakien arteko komunikazio elkarrekintzen funtzionamendua zehazten duten faktoreak zorrotz ezarri nahi lituzke, ikuspegi zabala hartuta hiztuna eta bere ingurua hartzen ditu aintzat [Escandell, 2004].

Horrela, esan genezake semantikotik baino ikuspegi pragmatikotik ekin diogula gure eginkizunari. Badago, jakina, testuinguruaren araberako interpretazioa alde batera utzita, hertsiki informazio linguistikoa kontuan izanik espresio konplexuen esanahia ikertzea posible dela aldarrikatzen duenik semantika formalean. Kontuak kontu, informazio estralinguistikoa, alegia, adierazpen konplexuen interpretazioa pragmatikak ikertzen du eta eremu horretan aritu garela esan liteke. Izan ere, bertso bat ulertzea ez baita bertako puntu guztien esanahia banan-banan ulertzea bakarrik, badago goragoko asmo edo egoerarekiko erreferentziaren adierazpenik ere bertsolarien jardunean eta, hain zuzen ere, gako horien bila aritu gara. Gure ustez, bertso-testuak ez dira esaldi-

sekuentzia linealak, irizpide bati darraiote eta hori atzeman asmo genuen. Agur-bertsoetako diskurtsoaren nondik norakoak ikertzeak erreferentziei so egitea zekarren ezinbestean. Gainera, ahozko jardun horretan, txapelketan nahiz bestelako egoera lasaiagoan, aurrez zedarritutako mundu bat eta arau batzuk onartzen dira. Berebiziko garrantzia du testuinguruak halako komunikazio egoeratan adierazten dena ongi ulertzeko. Balizko araubide horren, aurrez ezagututzat jotzen dugunaren eta igorleak linguistikoki esaten duenaren arteko zubia erreferentziak aztertuz egin nahi izan dugu.

Pragmatika hizkuntzalaritzaren atal berri eta trebatugabea izatetik bere lekua eta besteen artean garrantzia hartzera aldatu zen 1970eko hamarkadan. Garai hartakoak dira egun ere egiten diren hainbat jardunaldi eta argitaratzen diren aldizkariak. Haatik, azken urteotan berebiziko garrantzia ari da bereganatzen hizkuntzalaritza konputazionalaren alorrean. Iker-arlo horretako hatsarre nagusien artean hizketa-egintzen teoria [Searle, 1969] [Austin, 1975] [Vidal, 2004] aipatu behar da. Esanahiaren arazoei konponbidea emateko funtsezko kontzeptuok baliatu zituen, asmoa, xedea edo intentzioa eta ekintza edo egintza (lokutiboa, ilokutiboa, perlokutiboa). Esan daiteke, aurrez ikusi bezala, gramatikak egiatasun frogen arabera aztertzen dituela esaldiak eta egia ala gezurra den esaterik ez badago inolako proposiziorik ez daraman enuntziatutzat dituela. Austinek delako esaldiok zentzua izan bazutela frogatu zuen. Gure lanerako erabilgarria izan zaigun beste ideia bat ere aipatzen du, konbentzionaltasun mailena: aztura soziokulturalei, gizarte egoerari eta botere harremanei zor zaien gertaera. Alor honetan aipagarria da orobat, Sperber eta Winsonen [1987] garrantziaren edo egokitasunaren teoria. Aurrez aipatu hizketa egintzen teorian du abiapuntua, baina ezagutza-mekanismoen funtzionamendua mezua igortzeko garaian aztertzen dute eta baita enuntziatuen interpretazioa ere. Gizakiak inferentzia bidez interpretatzen ditu enuntziatuak, ez ditu balizko interpretazio guztiak aintzat hartzen; aldiz, eskuragarri dagoena eta prozesatzen errazena dena aukeratzen du, igorri duenak ere hori egingo zuelakoan baitago.

Komunikazio prozesuetan zein ote da testuinguruaren egitekoa, papera. Verschueren-ek alor honetan egindako lanean [1999] jasotzen duen bezala hiru alderdi nagusitan bereiz daiteke testuinguruko informazioa: a) Batetik, partaideak leudeke. Beraien egitekoa zehaztu beharko litzateke eta baita zein ote den haien arteko indar-

harremana. b) Bestetik, mezuaren edukia ere aztergai garrantzitsua da; zenbateraino ote da onuragarria nahiz kaltegarria hartzailearentzat? c) Azkenik, komunikazio ekintza bera definitzea proposatzen da: ea elkarrizketa, irakurketa, mediku txostena edo beste zernahi ote den.

2.2.1. Diskurtsoaren analisia

Hizkuntzalaritzaren iker-adar honetatik ere begira dakioke eskuartean dugun eginkizunari, beti ere, asmo dugun jomuga erdieste aldera. Diskurtsoaren analisian dihardutenek perpausaren esanahitik haratagoko nozioa behatzen dute. Enuntziatu edo esapide deritzona da lanerako gaia, hots, testuinguru jakin batean idatzi edo esan diren esapideak aztertzen dituzte. Horretarako testuingurua berariaz begiztatu eta ondoko galderen erantzunaren xerka aritzen dira:

- Zein dira diskurtsoan parte hartzen dutenak? Zein harreman dute? Ezagutza maila ezberdina ote dute? Zein helburu dute?

- Nola dakigu igorleak zer esan nahi duen? Zer esan nahi du hizkuntza pieza honek testuinguru honetan? Zer esan nahi du benetan igorleak? Zein faktorek ahalbidetzen du interpretazioa? Zer behar dugu testuingurua ezagutzeko? Zein dira argibideak, zantzuak, aztarnak?

Gu ere halaxe aritu gara hein batean eremuko lanean eta corpusa behatzen jardun dugunean. Benetako datu linguistikoak ere aztertu ditugu, testuinguruari zegozkionak lehenetsita ordea. Zentzu honetan aipagarria izan daiteke alderdi etnografikotik Hymes-ek [1972] hizketa ekintzatan aurkitu zuen elkarrekintzako osagaien taula:

Eszena antolatzen -->	egoera fisiko nahiz denborazko subjektiboak, zeintzuek uneko gertaera definitzen duten.
Partaideak -->	igorle/hartzaile/publiko/aipatuak/aipatuak entzuten dituena.
Amaierak -->	Asmoak eta helburuak, xedeak
Ekintza sekuentzia -->	mezuaren forma eta edukia.
Gakoa (kodea) -->	tonua, hitz egiteko modua.
Tresnak -->	Ahoz, keinuz eta abar.
Elkarrekintza arauak -->	Kulturalki zehaztutako ezaugarri berariazkoak
Generoa -->	testu kategoriak

1. taula: elkarrekintzetako osagaiak

Goiko osagaiak lan honetan jasotako esparru zedarritutik at dauden arren, aintzat hartu beharrekotzat jo ditugu. Haatik, finean Laboven aldaketa-teoriari [1972] atxikitako nahiz haren tankerako egituraren bat aurkitu nahi genuke agur-bertsoen egituratzat. Hark testu orotan eskuarki jarraitzen den hurrenkera kanonikoa proposatzen bazuen, gurea ataza zehatzagoari lotutako berezitasunei so osatuko genuke. Hona Labovek proposatutako egitura kanonikoa:

***Testuaren egitura**:* *Laburpena*
Orientazioa
Konplikazioa
Ebaluazioa
Ebazpena
Borobiltzea

Halliday-en eta Hassan-en ekarpenetan erreferentziaz gain lokailuak ere aipatzen dituzte testu multzoen arteko egitura linguistikoak aurkitzeko [Schmitt, 2002]. Ez gara honako ahaleginean halako ekimenetan saiatu. Lokailuen erabilerak garrantzi handiagoa luke testuaren tamaina bertso bakarrean sartzen diren hitz kopuru xumea biderkatzen duten laginetan.

Esan bezala, hitz, sintagma, esaldi eta perpausetatik haratagoko ezagutza ezinbestekoa da komunikazio arrakastatsua lortze aldera. Esaten duenarekin igorleak adierazi nahi duen eta nola lortzen duen hura ulertua izan dadin. Bestalde, hizkuntza, gizarte eta kultur testuinguruen arteko harremana ere azter liteke. Hizkuntzak munduaren ikuspegiak eta hura ulertzeko era berriak nola sortzen dituen.

2.2.2. Oinarrizko unitatearen eta kategoria-sistemaren bila

Erabaki beharra izan genuen aztergaitzat hartzeko unitatea hitza, puntua ala bertso osoa izango ote zen. Bertsoan zeharreko erabakien eta diskurtsoaren antolaeraren berri emango zigun egitura lortzekotan, puntua jo genuen ideiak adierazteko zein haiek behatzeko unitaterik egokientzat. Bertsoak puntuka zatituta haietan aurki genitzakeen ezaugarriak edo edukiak identifikatzea zen hurrengo ataza. Ahalik eta ezaugarritze zabalena egitea genuen helburu, aldi berean, puntuak ahal bezainbeste bereiziko dituen atributu zerrendarik zehatzena aukeratze aldera.

Bertsozale elkarteko corpusetik (hirugarren kapituluan deskribatuko dugu) 40 bertso lagin gisa hartu eta puntuz puntu aztertu genituen (212 puntu guztira).

Kategorizatzerakoan, berriz ere, bi aukera aurreikusten ziren. Batak, sumatutako gai guztiak aztertu eta multzotan bildu ostean sailkapenik egiazkoena proposatzeko modua eskaintzen zigun. Bigarren bideak, aldiz, hipotesi batean zuen oinarria eta lortutako datuak hara makurtzea edota lehen ustea datuetara egokitzea eskatzen zuen. Bietara jokatzea erabaki genuen.

Aurreneko gerturatze honetan, metodo induktiboa erabiltzen saiatu ginen. Laginekin taxututako bertso-puntuak aztertu eta haietan azaltzen ziren gai nagusi edo arrazoizkoenak zerrendetan jaso genituen.

Puntuka bereizitako bertsoetan antzematen genituen gaiak aukeratutako unitate bakoitzaren ondoan zerrendatu genituen. Gaiak aztertu eta zerrenda beregain bat egiteko saiakera hartu genuen jomugatzat. Azken zerrenda murriztu, murrizgarri eta egokitu beharrekoa osatzea erdietsi genuen, zeina behar bezala mailakatzen ere ahalegindu ginen.

Edukien zerrenda luzea lortu dugu, baita balizko kategoria sailkapen proposamenak egin ere haietatik abiatuta. 2. taula, saiakera haien adibide gisa eskaintzen da. Garrantziaren nahiz agerpen kopuruaren araberako hurrenkeran egituratutako kategoria taula.

Txapelketakoak				
Hasierakoak			**Bukaerakoak**	
Agur formalak Eskerrona. Entzuleak ahotan. Omenaldia bidegileei eta jarraitzaileei.	Saio ona eskaintzeko itxaropena. Norbere asmoak, ahalegin betea, erronkaren neurria. Estutasuna, larritasuna.	Epaileei. Finaleko kideei. Gai jartzaileari	Agur formalak	Norbere egoera
Saioa egiten den lekuari erreferentzia.	Apaltasuna. Bertsolaritzaren alde, goratzarrea.	Lanari buruzko azalpenak. Iruzkinak bertsolaritzaz eta txapelketaz.	Gizarte gaiak	Kritika soziala. Unean uneko gertaerei eta bistako elementuei erreferentzia.
Unean uneko elementuei erreferentzia. Albistegietako berriak, bolo-bolo dabiltzan kontuak.	Urte sasoiari dagozkionak.	Bertsolarien asmoak. Aurrekariak	Egindako lanari gainbegiratua. Txapelketako ibilbidea	Txapelaren eskaintza. Txapeldunari aitorpena Hurrengo saiokoei onena opa.
Gizarte gaiak	Bertsolaritzaren definizio bitarra. Zer behar du bertsolariak eta bertso on batek.	Ahalegina eskaini talde-lanean aritzeko.	Hurrengo saiorako asmoa	Saioa egin den lekuari erreferentzia
Etxekoak eta lagunak gogoan.	Elkartasuna, batasuna, Euskal Herria, Euskara, anaitasuna.	Norbere jatorria.	Txapelketari buruzko hausnarketa	Elkartasuna, batasuna, Euskal Herria, Euskara, anaitasuna.

2. taula: Bildutako gai zerrenda (abiapuntua)

Esan gabe doa, halako eduki zerrenda mardulak ez digula nahi genukeen bezala laguntzen gure ataza burutzen. Atributuen bilduma zedarritzen eta mugatzen saiatu beharra adostu genuen.

Corpus oparo hartan ausaz bildutako lagin haren barrenean puntuetan jarri dugu, esan bezala, begirada. Berrogei bertso horietako puntuak eskuz kategorizatzeari ekin genionean, hastapenetik ohartu ginen eremu lausoan genbiltzala, alegia, lana zehaztasunez burutzeko zailtasuna aipagarria zela. Hala ere, oinarriak finkatzen ari ginenez behar adina asti hartu genuen ataza osatzeko.

Eskuz kategorizatzen geniharduela erabaki horri sendotasuna eman asmoz, alorrean aritzen diren eta laguntzeko prest genituen bi adituri (Mikel Aizpurua[2] eta Karlos Aizpurua[3]) helarazi genien nola egitasmoaren informazioa, hala, puntuak sailkatzeko erreferenteei buruzko azken proposamena. Lehena, bertso-eskoletako irakasle izateaz gain, hainbat txapelketatako epaile gisa ere aritu izan da; gainera, zenbait liburu argitaratu ditu bertsolaritza gaitzat hartuta (biografiak, bertso hautatuak, ikas-materiala, transkribaketa irizpideak eta abar). Karlos berriz, bertso-eskola irakasle, bertsolari, gai-jartzaile eta oro har eragile gisa aritzen da. Hainbat argitalpen ditu bertsolaritzaren alorrekoak (aldizkariak, bildumak, gidoiak eta beste hainbat). Oro har, adostasuna erakutsi zuten egindako sailkapen ahaleginarekin. Sailkapen horri zein egiten ari ginen lanari egindako iradokizunak gogoan hartu eta aurrerago baliatzekotan gara. Hain zuzen ere, ondorioetan aipatuko ditugu iradokizun haietariko zenbait.

Alabaina, adituengana gure irizpenekin gerturatu ginen. Horretarako, diskurtsoaren azterketan, testu klasikoen sailkapenetan eta inprobisazioaren zein herri poesiaren alorretan egin dugu arrantza gure sailkapen induktiboa teoriatik abiatuko den prozedura deduktiboarekin uztartu eta sailkapen sendo bat aurkezteko.

Batik bat bost sailkapen modu, eredu, ideia edo irizpide hartu ditugu kontutan; azkena, horiek guztiak gogoan geronek moldatutako hipotesia litzateke. Jarraian zerrendatu ditugu irizpideen iturriak, oinarriak eta kategoriak erarik sintetikoenean.

2 http://bdb.bertsozale.com/web/haitzondo/view/-Mikel-Aizpurua
3 http://bdb.bertsozale.com/web/haitzondo/view/-Karlos-Aizpurua

2.2.2.1. **Alexis Díaz-Pimienta, *Inprobisazio poetikoaren barne dinamikaz*[4]**

Elkarrizketa antolatzeko baliabidea den deixia edo erreferenteen araberako sailkapena erabil genezake atributuentzako balioak bilatzeko.

Hauek dira bertsolariak erabili ohi dituen elementu erreferenteak:

a) Egoerazko deixia

(1) Pertsonazkoak (ni, zu, gu, zuek, entzuleak (*zu* kolektiboa))

(2) Espaziozkoak (hemen, leku honetan) Oso ohikoa hasierako eta amaierako agurretan

(3) Denborazkoak (gaur, orain, data zehatza, aurten, hil honetan, denborazko erreferente guztiak). Oso ohikoa hasierako eta amaierako agurretan

b) Testuinguruaren deixia

(1) Aurrez aipatutako hitzei eta gaiei egindako erreferentziak. Uneko gai sozialak, saioan aipatutakoak eta abar.

(2) Gai orokorrak, betierekoak (elkartasuna, batasuna, bertsolaritza, euskara eta beste)

2.2.2.2. **Guillermo Velázquez, *inprobisazioa*[5]**

Artikulu honetan nabarmentzen denez, Hupango menditarrek beti antzeko gaiak erabiltzen dituzte bat-batean aritzeko. Haiek jarraitzen duten gaikako bide antolaera geure erara moldatua.

a) Diosala egin, ohoratu, eskertu, agurtu
b) Betegarria
c) Ingurukoei eta haien esanei erreferentzia

2.2.2.3. **Brahim Baouch, *Poesia benetako balioen elementu bultzatzaile gisa*[6]**

Tamazigh hizkuntzan diharduten bat-bateko sortzaileek gehien darabiltzaten gaiak edo hizketa xedeak.

a) Bere gizarteko balioak sustatzea.
b) Herriarekiko maitasuna.
c) Laguntasuna (gizakien arteko harremanen oinarri leialtasuna).
d) Herrimina
e) Erresistentzia

4 In *Ahozko inprobisazioa munduan topaketak*. Donostia. 2004. Euskal Herriko Bertsozale Elkartea. 63-114 or.
5 In *Ahozko inprobisazioa munduan topaketak.* Donostia. 2004. Euskal Herriko Bertsozale Elkartea. 125-136 or.
6 In *Ahozko inprobisazioa munduan topaketak.* Donostia. 2004. Euskal Herriko Bertsozale Elkartea. 245-256 or.

2.2.2.4. **Aristoteles *Erretorika***

Diskurtsoaren egitura aztertzea xedetzat duen lan honetan ezin irakurtzeke eta aintzat hartzeke utzi klasikoak. Haien artean jargoian jar genezake Aristoteles eta *Erretorika* haren lana; edozein diskurtso antolatzerakoan jarraitu beharreko ibilbide-orria eskaintzen duena.

Prestakuntza urratsak	Antolaera	Helburuak	Erreferenteak
Argudio egokiak bilatu	Sarrera	Informatzea	Hiztunaren egoera
Ordena egokian jarri	Azalpena	Heztea	Entzulea, helburua
Egoki formulatu	Arrazoibidea	Pertsuaditzea	Ingurunea
Gogoangarri egin	Amaiera	Entretenitzea	
Diskurtsoa gauzatu			
Argumentuak	Ethos (igorleari begira, norbera)		
	Pathos (hartzaileari begira, entzulea)		
	Logos (gaiari begira, argudiaketa)		

3. taula: diskurtsoaren egitura (*Erretorika*)

Egile klasiko honen lanean oinarritzen da, orobat, Joxerra Garziaren *Jendaurrean hizlari* (2008) saiakera. Bertan sakon bezain zabal garatzen da bertsoetako diskurtsoaren eta erretorika orokorraren arteko uztarketa.

2.2.3. **Agur-bertsoetako puntuetan antzemateko atributuak**

Dagoeneko aipatu da puntuak ezaugarritzeko aurreiritzirik gabe ekin geniola lanari. Lagin-bertsoetako puntuak bereizten zituen ideia edo erreferente behinenak aukeratzen genituen 212 puntu haietako bakoitzari esleitzeko. Ondoren lortutako etiketak multzokatzen saiatu ginela ere azaldu dugu, haren emaitza da 2. taula.

Haatik, bagenuen aurreusterik, eta azken ataletan horiek sendotzeko alorrean diharduten hainbat egilek behaketa sakonen ostean eskainitako atributu edo ezaugarri sortak baliatu genituen (2.2.2.1-etik 2.2.2.4-ra arteko ezaugarri sailkapenak).

Xendra deduktiboa induktiboarekin uztartu eta aurrez bildutako gai zerrenda joria oinarrizko sei kategoriatan bildu genuen. Horrekin, agur-bertsoetako punturik gehienak etiketatzeko, eta halaber, zehaztasun maila nahikoa esanguratsua lortzeko gai izango ginen.

Behin-betiko kategoria edo ezaugarri zerrenda.

1.	**Mezua edo ideia nagusia**
2.	**Lekua**
3.	**Publikoa**
4.	**Saioa**
5.	**Norbera (asmoa, egoera)**
6.	**Betelana (oso bakana, baztergarria, behin eta berriz erabilgarria)**

4. taula: behin-betiko ezaugarri zerrenda

Egindako bidea erakuste aldera, 5. taulan bi bertso eskaintzen ditugu adibide gisa. Puntuen eskuineko lehen zutabean hastapenean hipotesiek estutu gabe eta askatasunez egindako etiketatzea genuke (2. taula osatzeko erabili genuena). Haren ondoan, berriz, bi bideen uztarketatik adostutako etiketak jarri ditugu.

Informazioa	**Puntuak**	**Hastapeneko ezaugarriak**	**Behin betikoak**
Millan Telleria 1986-03-09 Bertsolarien txapelketa nagusia. Tolosa Arratsaldeko saioa Hasierako agurra	Talde hontako sei lehen galbaiak hazitako utzi ginun,	Txapelketako ibilbideari erreferentzia.	Saioa
	ta gure zai zeuden biak ere ez datozte oso ilun,	Finaleko kideei.	Saioa
	ta entzuleak dana gainezka kabitu ezinik inun,	Publikoari.	Publikoa
	eta aurrean gu epaitzeko hona bederatzi lagun;	Epaileei.	Saioa Norbera (egoera)
	ia guztiak portatzen geran alkar gozatu dezagun. (bis)	Saio biribila osatzeko itxaropena.	Norbera (asmoa) Mezua (ideia nagusia)
Anjel Mari Peñagarikano 1986-03-09 Bertsolarien txapelketa nagusia. Tolosa Arratsaldeko saioa Hasierako agurra	Estu ta urduri nago ez trankil ta lasai,	Norbere egoera azaltzen. Estutasuna.	Norbera (egoera)
	lanean hasterako noiz amaituko zai,	Ez dago nahi lukeen bezala, zerbait gertatu zaio eta denek dakite zer.	Norbera (egoera) Saioa
	bertsozaleek berriz saio polita nahi,	Publikoari erreferentzia.	Publikoa Saioa
	barkatu ni ez naiz gaur horretarako gai,	Saio ona osatzeko itxaropena.	Norbera (egoera)
	asko ez det egingo bainan al dana bai.		Norbera (egoera) Saioa Mezua (ideia nagusia)

5. taula: etiketatze adibidea

2.3. Ikasketa automatikoa

Testuen kategorizazio automatikoa, hots, testu-dokumentuei bertan antzematen den edukiaren arabera aurrez definitutako kategoriak esleitzea, lanerako modu eta ikerketa esparru garrantzitsua da. Izan ere, egunerokotasunean eskuartean darabilgun testu-dokumentu kopurua eskerga baita. Auzi honetan nagusiki jarraitzen den hurbiltze modua ikasketa automatikoko (IA) metodoetan oinarritzen da. Delako metodootan sailkatzaileek automatikoki ikasten dituzte kategorien ezaugarriak aurrez sailkatutako testu sorta batetik abiatuta [Sebastiani, 2002]. Dokumentu sailkatzaile bat eraikitzeko ataza ez da bereziki aldentzen IAn baliatzen diren beste ataza moduetatik. Zentzu honetan, asko dira literaturan aurki daitezkeen era honetako hurbiltze proposamenak. Cardoso-Cachopok eta Oliveirak [2003] idatzitakoei jarraiki, dokumentuak irudikatzeko moduetan eta dokumentu hauetako bakoitza kategoria egokiari esleitzeko eran legoke bereizketaren muina. Beraz, urrats biek ala biek, alegia, dokumentuak adierazteko eta sailkatzeko moduak berebiziko garrantzia dute azken emaitza arrakastatsua izan dadin. Berariazko ataza bat burutzeko egokiagoa izan daiteke hurbilpen jakin bat egitea, hots, eskura izan ditzakegun datuei begira; alabaina, bestelako oinarriekin eta testuinguru ezberdinean baliteke hurbilpen hori ez izatea aproposena [Zelaia et al., 2005], [Kim et al., 2002], [Joachims, 1998].

2.3.1. Testu kategorizazioa

Testu kategorizazio metodoen xedea, dokumentu bati aurrez definitutako kategoria sail bat edo gehiago esleitzea da. Testu sailkapenaren berri damaigun azterketa aurkitu dugu Sebastianiren artikulu [Sebastiani, 2002] aipagarri batean. Ikerlarien artean luze-zabal onartua dago, dokumentuak irudikatzeko moduak berebiziko eragina duela sailkapenean erdietsitako emaitzen kalitate orokorrean [Leopold and Kindermann, 2002]. Gehienetan, dokumentu bakoitza hitzak biltzen dituzten array[7] gisa adierazten da. Trebatze atazetan erabili beharreko dokumentuetan bildutako hitz multzo osoari hiztegi edo lexiko esaten zaio. Hortaz, dokumentu oro bektore gisa ordezka daiteke. Osagai bat hiztegian

7 Hainbat alorretan erabiltzen bada ere, hizkuntzalaritza konputazionalean honela definitzen da: elkarrekin erlazionatutako mota bereko aldagaien multzoa. Hauek erreferentziatzeko izen bakarra eta indize bat erabiltzen dira.

dagokeen termino bakoitzari egokituko zaio, eta orobat, hitz horren dokumentuan zehar duen agerpen kopurua adieraziko duen zenbaki bat jarriko zaio ondoan (zero balioa hartuko du dokumentuan inoiz agertzen ez bada). Dokumentuak adierazteko edo errepresentatzeko era honi hitzen zakua edo *bag-of-words* esaten zaio. Corpusean izan daitezkeen terminoen zenbatekoa ikaragarria izan daiteke halako ereduak baliatzen ditugunean; hori da, hain justu, eredu honen alderik txarrena. Izan ere, ezaugarrien kopuru eskerga horrek baliagaitz egiten dute zenbait ikasketa-algoritmoendako. Hori dela eta, ezinbestekoa suertatzen da dimentsioak murrizteko metodoak baliatzea. Aipatu murrizketa egiteko bi modu erabili ohi dira: badago datuak aurrez prozesatzerik, esaterako, lema poltsako[8] hiztegiaren neurria kontrolatzeko zenbait iragazki ezar genitzake. Bestela, dimentsionalitatea murrizteko teknikak ere balia genitzake.

Dokumentuak kategorizatzeko ohiko prozesua bi urratsetan egiten da: a) *training* edo trebakuntza pausuan, prozesu orokor induktibo batek sailkatzaile bat sortzen du etiketatutako dokumentu sorta batetik ikasita. b) *Testing* edo proba aldia da bigarrena. Bertan sailkatzailearen errendimendua neurtzen da. Erabili dugun eskuz kategorizatutako corpusaren neurria handiegia ez izaki, *k-fold cross-validation* metodoa hartu dugu egokientzat. Tolesdura edo *fold* baliotzat K=10 baliatu dugu.

2.4. Erregelatan oinarritutako hurbilketa

2.4.1. Ezaugarri linguistikoak

Aurrerago zerrendatu eta azalduko dugu etiketatzeko eskema lortzeko osatu dugun ibilbidea, zeina adituen, alorreko literaturaren eta azterlarion irizpide zein irizpenen arabera osatu dugun. Kodetze, behatze, eztabaidatze eta berrikuste prozesu iteratiboa gauzatu da, corpuseko datuetan ikusten diren kontzeptu esanguratsuen adierazleak era induktiboan ezagutzeko. Azkenik, adituek egitura sendoa lortzeko bidea eskaini digute.

Aurreneko hurbilketan, gizakiak garatutako hizkuntza naturala prozesatzeko erregelak sortu eta ezarri ditugu etiketatutako bertso-puntuak erauzte aldera. Hurbilketa hau ezagutzan oinarritutakoa da, ezaugarri linguistikoak aztertzen ditu diskurtsoaren informazioa etiketatzeko, gure kasuan batik bat ezaugarri morfologikoak. Lan honetarako, hizkuntza naturala prozesatzen aditua den aztertzaileak erregelak proposatu

8 Hitzen lemek soilik osatzen duten *bag-of-words.*

eta egiaztatu ditu. Erregelak sortzeko, IXA taldearen analizatzaile morfo-sintaktikoa erabili genuen (EUSTAGGER[9]). Aztertzaile gisa EDBLren[10] kodeak gainbegiratu genituen, balizko egiturak goitik behera eta sakon ulertze aldera. Gainera, markatutako data berrikusi genuen kodeak testuan nola ezarri eta interpretatu direnaren behetik gorako ulermena izateko. Erregelak idazteko prozesua iteratiboa izan zen, zeinaren bidez erregelak testuetako kode adibiderik begien bistakoen eta ugarienak jaso asmoz idatzi ziren. Halaber, estaldura zein doitasuna uztartu nahi zirenez etengabe birfindu ziren lortutako erregelok.

Bi abantaila eskaintzen ditu erregelatan oinarritutako hurbilketak. Lehena, erregelatan oinarritutako saiakerari ez dio ezinbestean eragiten eskuragarri dagoen adibide kopuruak. Izan ere, giza adituek garatzen baitituzte eta aditu gisa duten ezagutza balia baitezakete. Bigarrenik, adituak badu datuen izaeraren arabera eta hura aintzat hartuta erregelak egokitzerik. Adibidez, ortografia eta gramatika akatsen eragina orekatzeko. Hala ere, saiakera modu honen kostua nahikoa handia da. Trebatutako profesional baten ahalegina eskatzen du, batik bat aztertu beharreko datuen zenbateko eskergari begiratu gero.

Ezagutzaren errepresentazioa edo datuen irudikatzea erregela multzo itxuran egin da. Lan honek badu onurarik, baita sistema aditurik sortu nahi ez denean ere. Datuak erregela bidez irudikatze horrekin fintze lanetan jardun baikaitezke, esan nahi baita, ikasketa automatikoko algoritmoekin trebakuntzan aritu ostean lortutako emaitzen arabera egokituta.

Hainbat eratako datuz osatutako corpus zabalak aztertzeak ikerlarientzako ahalegin handia dakar. Ikasketa automatiko hutsa eta erregelatan oinarritutako bideak erabiltzeak asko laguntzen dio ikertzaileari corpus eskergak aztertu nahi dituenean. Alabaina, benetako datuekin lanean diharduguneanean, giza etiketatzailearen beharra ezin ekidin daitekeela uste dugu. Oinarrizko kodetze lanetan eta hastapeneko erregela sorta bat osatu bitarte, behinik behin. Adituok ondoren automatikoki lortutako datuak fintzera bidera dezakete beraien ahalegina, doitasunari eta zehaztasunari eskainiz beraien arreta nagusiki.

9 http://ixa2.si.ehu.es/demo/analisimorf.jsp

10 4.1 atalean bi baliabideon deskribapen azkarra egiten dugu.

3. Esperimentu-saioak

3.1. Ikasketa automatikoko tekniken sailkapena

Ikasketa automatikoko teknikak modu askotara sailka daitezke, nahiz eta multzoen arteko mugak ez izan beti garbi-garbiak (sailkapenak egiterakoan maiz gertatzen den moduan, bestalde):

> Dependiendo del tipo de conocimiento a adquirir, podemos hablar de conocimiento (y aprendizaje) simbólico o subsimbólico. Desde el punto de vista de la forma del aprendizaje, se puede hablar de aprendizaje supervisado o aprendizaje no supervisado. Desde el punto de vista de las técnicas empleadas, podemos hablar de sistemas basados en técnicas estadísticas (o modelos estocásticos) y sistemas basados en razonamiento inductivo."}
>
> [Márquez, 2002]

Aipuko ikasketari bagagozkio, batik bat bi eratako ikasketa modua osa daiteke: **Gainbegiratua edota ez-gainbegiratuak.**

a) Eredu gainbegiratuan, klaseak finkatuta daude, eta ikasketa-adibide bakoitza zein klaseri dagokion jakin badaki ikasketa-algoritmoak. Helburua, beraz, orokortzea da, gerora adibide berriak (ikusi gabeak) sailkatzeko.

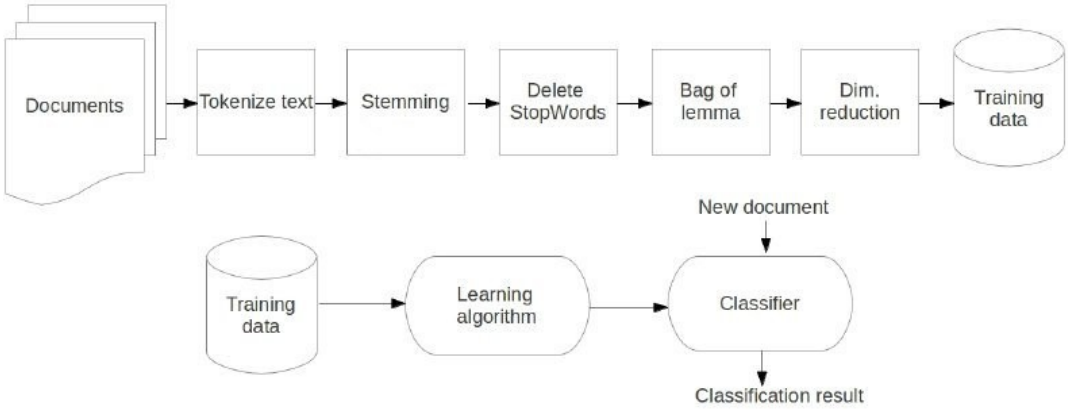
Batzuetan, ikasi behar dena inplizituki etiketatuta dator, baina, normalean, eskuz etiketatu behar da. Testuinguruaren araberako zuzenketa ortografikoan, esaterako, testuinguruaren arabera zuzenak ala okerrak izan daitezkeen hitzak testu zuzenetan zuzen idatzita daudela suposatzen da, eta ikasketa-adibide gisa erabiltzen dira (eskuzko etiketatzearen beharrik gabe). Gehienetan, ordea, ezin da halakorik egin, eta eskuz etiketatu behar izaten da ikasi nahi den kontzeptua.

b) Eredu ez-gainbegiratuan, berriz, klaseak ez dira ezagutzen aldez aurretik, eta ikasketa-algoritmoak gai izan behar du klase horiek zein diren ebazteko, antzeko adibideak multzotan (*cluster* deiturikoetan) bilduz.

3.2. Aurre-prozesatze lanak

Dokumentuak irudikatze aldera, hitz-zakuen ereduetako datuak aurrez prozesatu ditugu. Testuak sailkatze ataza burutzeko, dokumentuan agertzen diren hitz guztiak ez dira esanguratsuak, jakina. Eskuarki, aurrez prozesatze urratsa eman beharra izaten da

Corpusaren dimentsionalitatea murrizte aldera, eta orobat, datuak bateratzeko. Hala egiteak jakina, tresnen errendimendua areagotzen du.



1. irudia: kategorizatze prozesua

Aurre-prozesatze lanean baliatu ditugun teknikak hurrengo kapituluan azalduko dugu zehatzago. 1. irudian ikusten den eskemari so, esan dezagun *Bag-of-lemma* sortu arteko urratsak direla. Labur aipa dezagun lehena, hots, tokenizatzea. Testua token edo hitzetan banatzeko teknikari deitzen zaio horrela, eta batik bat hitz beregainak bereizteko prozesuaz ari gara.

a) Erro-bilaketa edo Stemming-a: erro bereko hitzak alboratzea ahalbidetzen du eta beraien artean ohikoena den elementua gordetzea. Euskararen inflexio morfologia dela eta, hitz jakin baten erroak hainbat egitura berri eskain ditzake. Euskararen deskribapen morfologiko laburra aurki daiteke [Alegria et al., 1996] lanean. Diogunaren adierazgarri edo adibidetzat balio beza *etxe* errotik eratorritako hitz zerrenda honek: *etxea*, *etxeak*, *etxeari* eta abar. Termino baten pisua zenbatesteko zeinahi hitz bere horretan eta zehatz hartuz gero, hitz horretatik eratorritako guztien antzekotasunak galtzeko bidea egiten ari garela esan nahi luke. Beraz, lematizatzaile bat erabili dugu. Euskararen deskribapen morfologikoan oinarritutako tresna honekin emandako ezein hitzen erroa bilatu dugu eta hura baliatu gure berariazko hiztegian [Ezeiza et al., 1998].

b) Lasto-hitzak (Stopwords): esanguratsuak ez diren hitzak ezabatu ditugu. Hala nola artikuluak, lokailuak eta zeinahi testutan maiztasun handia duten hitzak izanik haien

artean bereizteko aukerarik ematen ez digutenak. Bertsolaritzan gehien erabili ohi diren hitzen zerrenda erabili dugu delako lasto-hitzen bilduma osatzeko.

Dimentsio-murrizketa. Testuak sailkatzeko ahalegin askotan izaten da ohiko urratsa dimentsio-murrizketarena. Halakorik egiteak, eskura dugun atributu sorta bilduma eraldatu eta txikiago bat osatzea dakar, bide batez, zorionekoak ginateke zerrenda berriak aurre esate gaitasun handiagoa balu; izan ere, hori baitugu azken xedea. Hona dimentsio-murrizketa egiteko erabiltzen diren bi moduak:

a) Corpusaren dimentsionaltasuna murrizteko asmoari jarraiki, ezaugarri jakinak aukeratzen eta ezabatzen dituen jardunbidea daukagu. Erauzitako ezaugarriak sailkatze atazan lagungarri ez diren eta bereizle lanik betetzen ez dituztenak izaten dira [Forman, 2003]. Jarraian, metodorik entzutetsuenak eta usu erabiltzen direnak aipatu nahi izan ditugu: *Information Gain, Chi-square eta Gain Ratio* [Zipitria et al., 2012].

b) Bigarren jarraibidea ezaugarri eraldaketarena litzateke. Ildo horri jarraituz gero, atributuen jatorrizko zerrenda egokitu eta trinkoagoa den berria egituratzen da. Hona, lanean aintzat hartu dugun ezaugarri eraldaketa metodo parea: *Principal Component Analysis (PCA)* [Wold et al., 1987] eta *Latent Semantic Analysis (LSA)* [Deerwester et al., 1990] [Hofmann, 2001].

Aipatu berri ditugun bi hurbilketen arteko ezberdintasun nagusia nabaria da, esan nahi baita, ezaugarri aukeratze metodoetan jatorrizko atributu sortatik azpimultzo bat hautatzen da, aldiz, ezaugarri-eraldatze prozeduretan itxuraldatu egiten dira hasierako ezaugarriak eta multzo berria osatzen. Azken bide honi darraiogunean, gogoan izan behar dugu burututako ekintzak eragina izan dezakeela emaitzak ulertzeko moduan, izan ere, eraldatutako atributuek errendimendu ona eskain diezaguketen arren, interpretatzen zailak dira.

3.3. Ikasketa-algoritmoak

Ez dugu sailkatzaileak emaitza egokiak ematea bakarrik nahi, aitzitik, datuetatik sailkatzaile egokiak egiten ikasi nahi dugu. Badira kategoria ezberdineko kasuak, baina ezaugarri berekin, horrek ziurgabetasuna eragiten du. Ahalik eta ziurgabetasun gutxien duten sailkatzaileak sortu nahi ditugu.

Etiketatutako corpusa daukagunez, gerta daitekeena aurreikusteko erabili ohi diren **gainbegiratutako ikasketa sistemak** erabili ditugu, izan ere, handik lortzen ohi den...

- Erregela ulerterraza izan ohi da.
- Azaltzen duen gertaera baino sinpleagoa izan ohi da erregela.
- Salbuespenak ikusteko ere erabiltzen da.

Hainbat izan dira ikaste automatikoko teknikak erabiliz lortutako eta proposatutako testu-sailkatzaileak. Bada horien berri luze-zabala arloko literaturan [Sebastiani, 2002]. Nahiz eta, diogun bezala, aurkeztu eta egiaztatutako hurbilpenen kopurua eskerga izan, testu kategorizazioarena iker-arlo bizia da oraindik ere, batik bat, gaurdaino akatsik gabeko saiakerarik ezagutzen ez delako. Hona aurkezten dugun lanerako erabili ditugun algoritmoen zerrenda: *Nearest Neighbour Classifier (IBk)* [Dasarathy, 1991], *Naive Bayes Classifier (NB)* [Minsky, 1961], *J48 Decision Tree Learner* [Quinlan, 1993], *Bayes Net* eta *SMO Support Vector Machine* [Joachims, 1998]. Multzotan banatuko ditugu erabilitako algoritmoak, beraien familiak deskribatuta (K-nn, *Naive Bayes, erabaki-zuhaitzak eta support vector machine*).

3.3.1. K nearest neighbour (IBk)

K-nearest neighbour (K-nn) metodoaren erabilera aspaldikoa da sailkatze metodo gisa. Aztertzen ari garen auziaren gertuko beste baten ebazpena egiaztatu eta hura proposatzen du aztergaiaren kategoriatzat. Distantzia kalkulatzeko ez da kategoriarik hartzen kontuan. Aldagaien artean bada kualitatiborik ere, denak ez baitira zenbakiak. Konputazionalki garestia da, datu berri bakoitza matrizeko aldagai guztiekiko distantziaren arabera sailkatu behar baita. *K*-ren lekuan hainbat zenbaki jar daitezke, kontuan hartu nahi den distantziaren arabera. Gure kasuan distantzia 1era daudenak aintzat hartuta lortu dira daturik interesgarrienak.

3.3.2. Naive Bayes (NB)

Probabilitatearen banakako banaketan oinarritzen da. Adibide baten klasea asmatzeko, behatutako adibidearen probabilitatea maximizatzen duena aukeratzen da. Horretarako, Bayes-en teorematik eratorritako formula sinple bat erabiltzen da, non atributu guztiei

dagozkien balioak emanik (a_1, a_2,... a_3) emaitza-atributuaren klase probableena (V_{nb}) aukeratzen baita:

$$V_{nb} = argmax_{v_j \in V} P(v_j) \prod P(a_i|v_j)$$

Naive Bayes sailkatzaileak hipotesi batekin jokatzen du: atazaren deskribapenerako erabilitako atributu edo ezaugarri bakoitza beste edozein bezain garrantzitsua dela; alegia, independenteak direla atributu guztiak. Hipotesi hau, ordea, ez da betetzen askotan, baina, hala eta guztiz ere, baldintza hori betetzen dela suposatzeak dakarren sinplifikazioak eredu dotore eta eraginkorrak eman ohi ditu .

3.3.3. Erabaki-zuhaitzak (J48)

Ikasketa automatikoko eskema klasiko hau *zatitu eta irabazi* teknikan oinarritzen da, eta grafikoki adierazita, zuhaitz baten itxura hartzen du; hortik datorkio izena. Erabaki-zuhaitzak sortzeko prozesua modu errekurtsiboan azal daiteke. Lehendabizi, atributu edo ezaugarri bat aukeratu behar da erro-adabegian kokatzeko, eta bere balio posible bakoitzeko adar bat egiten da. Gero, prozesua errepika daiteke errekurtsiboki, adar bakoitzerako, baina adar bakoitzeko baldintzak bete dituzten adibideekin soilik. Adabegi-ume bakoitzeko adibide guztiek sailkapen bera dutenean amaitzen da prozesua, kasu horretan ezingo baita adabegia gehiagotan banatu; adabegi hori, beraz, hostoa izango da. Erabaki beharreko gauza bakarra, eskema honetan, zera da: une bakoitzean aukeratu beharreko atributua. Unean uneko atributuaren aukeraketak, ordea, berebiziko garrantzia dauka, behin atributu hori erabili eta gero ez baita gerora hartuko diren erabakietan atributu bera berriz erabiltzen. Bestalde, geroz eta atributu gehiago izan, orduan eta denbora gehiago beharko du ikasketa"-algoritmoak.

Eskema honen abantailetako bat, zera da: ikasketa sinbolikoko algoritmo bat izanik, eskuratutako ezagutza adierazpide ulergarri batean jar daitekeela; alegia, zuhaitz formako hierarkia bat osatzen dute erabaki-zuhaitzek, eta hierarkia horretatik erregelak erauz daitezke. Hala, gaian aditua denak erregela horiek interpreta litzake. Desabantailen artean, zenbakizko atributuak erabiltzeko dituen arazoak aipa daitezke. Izan ere, zenbakizko atributuak erabiltzeko moldaketaren bat egin beharra dago eskema hau erabiltzekotan; zenbaki bakoitza atributu bitar bihurtzea da moldaketarik ohikoenetakoa

(diskretizatzea[11], alegia). Askoz era naturalagoan egiten du lan zenbakizkoak ez diren atributuekin.

Hizkuntzaren prozesamenduko ia maila guztietan erabili izan dira erabaki-zuhaitzak. Emaitza onak lortu izan dira, esaterako, ahotsaren ezagutzan, morfosintaxiaren etiketatzean, desanbiguazio semantikoan, analisi sintaktikoan, laburpenen sorkuntzan, entitateen ezagutzan, dokumentuen sailkapenean eta itzulpen automatikoan [Márquez, 2002].

3.3.4. Support Vector Machines (SMO)

Ikasketa automatikoko algoritmo honek eredu linealek dituzten desabantailak konpontzen ditu. Izan ere, linealak ez diren datu multzoentzat soluzio bat ematen du. Hainbat atazatan erabiltzen da eredu lineal hau.

Support vector machines (SVM) izeneko ikasketa automatikoko eskema hau ez da batere azkarra; adibide asko dituen ikasketa corpusekin lan egiterakoan, batik bat. Gainera, ez da sinbolikoa; beraz, ezin da eskuratutako ezagutza gizakiarentzat ulergarria den adierazpide batera ekarri. Hala eta guztiz ere, emaitza onak lortzen ditu oro har, erabaki"-muga konplexuak eta finak eskuratzen dituelako, eta portaera bereziki ona dauka atributu askoko atazetan, eta baita linealki banagarriak ez diren problemetan ere.

3.4. Tf-idf (term frequency-inverse document frequency)

Bilduma baten barrenean hitz bat dokumentu batentzat zein garrantzitsua den balioesteko erabili ohi da maiztasun adierazpide hau.

Batetik, terminoaren maiztasuna (term frecuency) daukagu, hots, ezein hitzek zeinahi dokumentutan duen agerpen kopurua**.** Bestetik, dokumentuen alderantzizko maiztasun faktoreak (inverse document frecuency) bilduman askotan agertzen diren terminoen pisua murrizten du eta orobat gutxitan azaltzen direnen pisua areagotzen.

Esaterako, testu-dokumentu sorta bat daukagu eta *txapel hau zuentzat* txatalarentzat esanguratsuena zein den zehaztu nahi genuke. Hauek bereizte aldera, dokumentu bakoitzean aipatutako termino orok (*txapel*, *hau*, *zuentzat*) duen agerpen

11 *Aldagaien balio posibleak infinitu izan beharrean, balio jakin batzuk ezartzen zaizkio, adibidez: handia / txikia / oso txikia/*

kopurua batu behar dugu.

Baina *hau* determinatzailea hain da arrunta, ez baita dokumentu adierazgarriak eta esanguratsuak ez direnak bereizteko hitz-gako egokia. Alabaina, *txapel* eta *zuentzat* terminoak gutxitan agertzen direnez, egokiagoak izan daitezke dokumentu garrantzitsuak bereizte aldera. Beraz, IDF faktoreak *hau* terminoaren pisua murrizten du eta *txapel* zein *zuentzat* hitzenak areagotu.

Baliabide estatistiko hau maiz erabiltzen da informazioa atzitze prozeduratan eta corpusetako datuak biltze atazatan, hala nola bilatzaileetan, testu laburpen automatikoetan eta *lasto-hitzak* (*stop-words*) iragaztea eskatzen duten atazetan.

3.5. Multi-sailkatzaileak

Egoera edo klase guztietan emaitza onak edo are onenak lortzen dituen algoritmo bakarrik aurkitzen ez denean, hainbat teknika uztartzeko aukera hobesten da. Algoritmoak konbina daitezke eta baita dimentsio-murrizketa teknikak ere. Sailkatzaile bakunen emaitzak konbinatuta osatzen dira, helburutzat *zehaztasuna* hobetzea hartuta.

Iragan mendeko 90. hamarkadan indartu zen iker-arlo hau ikasketa automatikoaren barrenean. Ordutik gaurdaino, hainbat proposamen eta metodo egin izan dira, baina finean egoera nahiz klase ezberdinetan emaitzarik onenak lortzen dituztenak bateratzean legoke ahaleginaren arrakasta.

Era honetako sailkatzaileak bi multzotan bereizten dira. Algoritmo bera erabilita datu multzo anitz ezarriz gero **sailkatzaile homogeneo**ez dihardugu (*bagging*, *boosting* eta abar). Aldiz, datu multzoa (atributuak, ezaugarriak) eraldatu gabe hainbat ikasketa-algoritmo erabiliz gero **sailkatzaile heterogeneo**ekin ari gara (*vote* eta antzekoak).

Lehenengo multzokoekin saiatu gara gu. Eta beraien artean *bagging* metodoarekin egin dugu proba. Honek sailkatzaile bateratu eta bakunak sortzen ditu trebakuntza datu-sorta laginak elkarren artean konbinatuz. Korrelaziorik gabeko sailkatzaileen zehaztasuna hobetu ohi du eta eskuarki trebakuntza (*training*) lagin gutxirekin aritzeko egokia izaten da [Stefanowsky, 2008].

3.6. Esperimentuetan erabilitako tresna (WEKA)

Esperimentu denak iturburu irekiko WEKA inplementazioan edo tresna erabilita gauzatu ditugu [Hall et al., 2009]. Java hizkuntzan idatzitako programa da WEKA eta libre bezain doan eskura daiteke egitasmoaren webgunetik[12], GNU lizentzia publikoa baitu. Zeelanda berriko Waikato unibertsitateak garatu duen baliabidea ikasketa automatikoan zein datu-meatzaritzan erabiltzeko sortu dute.

Ikerkuntzan zein hezkuntzan egin du lekua eta ezaugarri nagusien artean aipagarriak dira: interfaze grafikoa eskaintzea, datuak aurrez prozesatzeko aukera, ikasketa-algoritmoak eskura jartzen ditu eta ebaluazio metodoak ere inplementatuta dauzka.

3.7. Corpusa

Askotariko bertsoekin (txapelketakoak, plazakoak, omenaldiak, bazkari nahiz afarietakoak eta beste) ahalik-eta corpusik osoena eratzea genuen lehen helburua. Bildu genituenen artean txapelketako bertsoek eskaini ziguten azterketarako multzorik egokiena, nola kopuruan hala banaketan ere (egileak, urteak, lekuak eta abar). Beraz, gure saiakera txapelketako aleetara mugatzea erabaki genuen. Ez dugu haien barrenean inongo bereizketarik egin, aitzitik puntu bakoitzean darabiltzaten gaiei egin diegu so, saioaren hasieran zein bukaeran eskainiak izan diren begiratu gabe.

Irizpideak argi, hurrengo urratsa corpuseko berrogei bertso biltzea zen. Eskuzko etiketatze lanetan gehiegizko pisurik hartu gabe, ahal bezain lagin esanguratsua lortzeko egokitzat jo genuen kopuru hori. Unitatetzat puntua genuenez, finean, 212 instantziako corpusarekin ekin genion behaketari.

Ahozko datuak izaki, hizkuntza estandarretik aldentzeko joera erraz aurreikus zitekeen. Hori aintzat hartuta eta gure estandarizatze-prozesuan garai jakinak daudela erabakita, laurogeiko hamarkadatik aurrerako bertso-saioak bakarrik jaso genituen. Bost urtean behingo tartean ausaz bildu ditugu estaldura ahal bezainbeste hedatze aldera. Izan ere, bertsolariekin batera garaiak ere aldatu egiten dira eta nola ez, baita ohiturak ere (tartean egon zitezkeen 1962an Mattinek Donostian kantatutakoa zein Sustrai Colinak

12 http://www.cs.waikato.ac.nz/ml/weka/

2009an Barakaldon botatakoa). Bilketa lana bertsozale elkartearen datu-basean osatu dugu eta haien lan eskerga bezain eskergarria baliatu (http://bdb.bertsozale.com/). Oinarrizko irizpideen artean jarri behar da, orobat, azterketa automatikoaren bidetik, ahalik eta eskuzko lan gutxien egitea hobetsi dugula eta transkribatutako bertsoen aukera hori baliatzea halabeharrezkoa iruditu zaigu.

Aurreprozesatze prozesua azaldu dugunean (3.2 atala) aipatu dugu dagoeneko *stemming* prozedura jarraitu dugula lortutako hitz guztien artean erroak soilik biltzeko, hots, *token* edo elementu esanguratsuak bahetze aldera. Hori eginda, lematizatzaileak 851 atributuko multzoa eskaini zigun. Nolanahi ere, goragoko lerroetan aipatu dugunez, estandarizatzeko ahalegin bat ere egin behar izan dugu, erro bera izanik ere ahoskeraren araberako transkribatzeak sor zitzakeen oztopoak gainditzeko. Prozesu horretan, XUXEN zuzentzaile ortografikoaren laguntza izan dugu eta eskuzko estandarizatze lanaren ostean, 614ra jaitsi zen atributu kopurua.

Zenbat eta atributu gehiago aintzat hartu behar, orduan eta zailagoa da sailkatzaileentzat *klase* egokian esleitzea. Hori dela eta lasto-hitzak biltzen zituen dokumentua iragazki gisa erabili eta gure lagina findutako 582 atribututan errenditu genuen.

Emaitzen deskribapenean zein ondorioetan ere adieraziko dugunez, esperimentuek aurrera egin ahala sailkatzaile hoberentzat genuenari on egingo ziola jakinda, lagina 15 bertso gehiagorekin handitu genuen. Tauletan bereizita eskaintzen dugu haiekin lortutako emaitza zerrenda.

Aurre-prozesuaren barrenean sartu dugu, halaber, *dimentsio-murrizketa* (3.2 atala). Teknika horiek ezarri ondoren % 10eko murrizte tasa adostu genuen, iker-arloko hainbat lanetan egindako proposamenei jarraiki.

3.8. Ebaluaziorako neurriak

Ebaluazioari dagokionez, sailkatzaile automatikoak ebaluatzeko PARSEVAL neurri hauek [Black, 1991] erabili ohi izan dira: doitasuna (*precision*) eta estaldura (*recall*). Hartutako erabakien zuzentasuna neurtzen du doitasunak; estaldurak, berriz, zuzenak direnetatik asmatzen direnen portzentajea ematen du. Gurearen antzeko lanetan eskuz

etiketatutako osagaiak hartzen dira zuzentzat. Horrela, sailkatzaile automatiko batek bertso-puntu bat zuzen etiketatu duela esango dugu, baldin eskuz etiketatutako osagaiaren klasea (guk erabakitako 6etako bat) bera bada.

Aipatutako neurriak 6. taularen gisako baliabidean oinarrituta kalkulatzen dira. Gure kasuan, *bai* ala *ez* zen aukera (0 ala 1).

a	b	<-- classified as
100	44	a = 0
47	21	b = 1

6. taula: confussion matrix edo kontingentzia taula

"100" zenbakiak benetan aztergai genuen klasekoak izanik, etiketa hori bera esleitu zaien instantzien kopurua ematen du. "44" zenbakiak egiaz behatutako klasekoak izanda ezezkotzat jo dituen puntu kopurua adierazten du; "47" zenbakiak, esaterako *Lekua* klasekoak diren, baina hala aukeratu ez diren instantzia kopurua adierazten du; azkenik, "21"' zenbakia klase jakin batekoak izanik, sailkatzaileak balekotzat hartu dituen puntuen kopurua da.

Doitasuna eta estaldura emaitzaren klase posible bakoitzeko kalkulatzen dira. Oro har, honela definitzen dira bi neurri hauek hizkuntzaren prozesamenduan, sailkatzaileak neurtzen gabiltzanean (A_z analizatzaile automatikoak zuzen etiketatutako osagai kopurua izanik; A_e analizatzaile automatikoak etiketatutako osagai kopurua izanik; E_e eskuz etiketatutako osagai kopurua (zuzentzat hartutakoak) izanik):

$$\textit{Doitasuna} = A_z/A_e$$

$$\textit{Estaldura} = A_z/E_z$$

Bagging multi-sailkatzailearekin (1-nn, None, 80[13]) lortutako datuak ziren adibidekoak. Hara zein emaitza islatzen duten, datu errealon karietara:

Doitasuna --> 100/100+44 = 0,694

Estaldura --> 100/100+47 = 0,680

Klasea benetan guk proposatutako seietako bat ez denean emaitza horiek lortzen

13 K-nn distantzia bakarrera, algoritmo gisa, dimentsio-murrizketarik gabe eta 80 atributu hautatuta.

dira. Baiezko kasuetan, berriz, honelakoak emango lituzke.

Doitasuna --> 21/21+47 = 0,309

Estaldura --> 21/21+44 = 0,323

Ez dago doitasunaren eta estalduraren artean erlazio matematiko zuzenik, baina sistemak detektatutako elementuen kopurua handitzen bada, haiek zuzen etiketatzeko aukera murriztu egiten da eta alderantziz. Ez da, beraz, erraza bi neurriok batera parekatzea. Hori dela eta, biak aintzat hartzen dituen zenbait neurri proposatu dira. Maiz erabiltzen den *F measure* edo *neurria* baliatu dugu hemen [Arrieta, 2010].

$$F_1 = \frac{2 * Doitasuna * Estaldura}{(Doitasuna + Estaldura)}$$

Horrez gain, zehaztasuna edo *accuracy* izeneko neurria ere erabiltzen da: hartutako erabaki guztietatik zuzenak izan direnen portzentajea neurtzen du. Kontingentzia-taulako (*confussion matrix*) datuekin kalkulatzen da neurri hau ere. 6. taulako adibideen arabera hortaz:

Zehaztasuna = (100 + 21) / (100 + 44 + 47 + 21) = 0,571

Emaitzak taula bidez deskribatzerakoan F measure-z gain zuzen sailkatutako instantzien portzentajea ere eskaini dugu (*performance*).

3.9. Emaitzen deskribapena

Esperimentuetan lortutako emaitzen berri ematen da atal honetan. Sailkatzailearen doitasuna adierazteko zenbait taula eskaini ditugu, zeintzuetan asmatze tasa balioesteko egokientzat jotzen diren neurriak eskaintzen diren (aurreko atalean azalduak).

Kontingentzia-taula ere erabili dugu zenbaitetan, egin ere, gainbegiratutako ikasketa saiakeratan emaitzak bistaratzeko beste modu egokia delako eta informazio osoaren zenbait alderdi behatzen laguntzen zigulako.

Dimentsio-murrizketa bitartekoak edo baliabideak bitan banatu ditugu 3.2 atalean azaldu bezala (atributu aukeraketa dagitenak eta emaitza gisa atributu konbinaketa eskaintzen dutenak). Tauletan ere nabarmen utzi nahi izan dugu ahal izan den neurrian.

Lehendabiziko taula honetan *mezua* gisa izendatu dugun klaseari dagozkion datuak biltzen dira.

MEZUA		**Ebaluazio neurria**	**Dimentsio-murrizketa**						
			Atributu aukeraketa				**Atributu konbinaketa**		
			Bat ere ez	**InfoGain**	**GainRatio**	**ChiSquare**	**PCA**	**LSA**	**LSA+15**
Algoritmo sailkatzaileak	**1-nn**	Performance	% 64,6226	% 55,6604	% 65,0943	% 57,5472	% 45,7547	% 64,6226	% 68,3849
		F-measure	% 62,3	% 53	% 58,1	% 54,6	% 46,9	% 57,8	% 62,7
	5-nn	Performance	% 65,566	% 67,4528	% 67,9245	% 67,9245	% 44,3396	% 64,6226	% 68,3849
		F-measure	% 58,4	% 54,7	% 55	% 55	% 43,3	% 57,8	% 62,7
	BN	Performance	% 64,6226	% 61,3208	% 65,0943	% 64,6226	x	% 67,9245	% 72,8522
		F-measure	% 57,2	% 54,1	% 53,6	% 56,1	x	x	x
	J48	Performance	% 65,566	% 65,566	% 65,566	% 64,6226	% 61,7925	% 67,9245	% 72,8522
		F-measure	% 57,2	% 56,6	% 56	% 56,1	% 61,4	% 55	% 61,4
	NB	Performance	% 67,9245	% 66,0377	% 66,9811	% 67,9245	% 51,4151	% 67,9245	% 72,8522
		F-measure	% 55	% 54	% 54,5	% 55,8	% 52,6	x	x
	SMO	Performance	% 61,3208	% 64,1509	% 59,9057	% 65,0943	% 59,9057	% 67,9245	% 72,8522
		F-measure	% 59,5	% 56,9	% 57,9	% 58,6	% 57	% 55	% 61,4

7. taula: 1. kategoria/klasea (***MEZUA***)

Datuen analisi azkar batek emaitzak benetan egokiak direla esatera eramaten

gaituzte. Izan ere, ehuneko 50aren gainetik azaltzen baitira sailkatzaile gehienak. Alabaina, ezinbestekoa zaigu ondoren eta emaitza guztien deskribapenean errepikatu bide den gertaeraren berri ematea. Horretarako erabiliko dugu, arestian azaldutako *confusion matrix* bat, hain zuzen ere, dimentsio-murrizketarik gabe *1-nn* algoritmo sailkatzaileak lortutako datuena:

a	b	<-- classified as
118	26	a = 0
49	19	b = 1

8. taula: *MEZUA* klasearen *confussion matrix* edo kontingentzia-tauletariko bat.

Aztergai genituen 212 instantzietatik 68 soilik dira guk aurrez mezua etiketarekin adieraziak. Proportzioan beraz, ugariagoak dira ezezko kasuak eta asmatze-tasa puztuta agertzen zaigu hori dela eta. Egiaz, gure xedeetarako egoki etiketatutako puntuak 19 baino ez liratekeela ikusten da nahaste matrizean.

Ezinbestean argitu beharreko puntu hori gogoan izanik ere, bada zer gehiago aztertzerik sailkatzaile egokienaren bilaketa honetan. Dimentsio-murrizketarik gabe, hots, instantziak dauden daudenean hartuta *Naive Bayes* algoritmoak eman dizkigu daturik onenak zuzen sailkatutako instantzien ehunekoari dagokionean, baina arloan neurri ziurragotzat jotzen den *F-measure* datuei so *K-nn* algoritmoak damaizkigu emaitzarik aipagarrienak (% 62,3), asmatze tasarik zinezkoenekin.

Nolanahi ere, sailkatzaileen lana asko hobetu zitekeela nabaria zenez, instantzien balioa neurtzen, ponderatzen eta orobat, aukeratzen saiatu beharra ikusi genuen. Finean, hautatutako algoritmoek emaitza esanguratsuagoak eskain zitzaten nahi genuen. Horretarako, gorago aipatu ditugun metodoak hartu ditugu kontuan (goreneko lerroan zerrendatuak), helburu horretan eskuarki erabiltzen direnak eta arloko ikerlanetan maiz ikusitakoak. Probatu ditugun tresnek ez dute egiazko asmatze kopurua areagotu. Askotan, haiek erabili gabe emaitza hobeak erdiesten dituzte sailkatzaileek, nahiz eta balizko doitasun portzentajeari begira itxurazkoa baino ez zen hobekuntza erakutsi. *LSA* (*Latent Semantic Analysis*) teknikak eskaini digu emaitzak hobetze bidean sendotasun itxura, esan nahi baita, algoritmo gehienetan eskaini ditu dimentsioa murrizteko beste

teknikekin findutako saiakeretan baino ondorio hobeak.

Itxaropen horrekin heldu genion azken zutabeko emaitzetan jasotako bide berriari. Jakinik *LSA*k terminoen arteko antzekotasun semantikoak finkatzen dituela eta haiek ezartzeko instantzia ugari aztertzea komenigarria zela ohartuta, 15 bertso berri erantsi genizkion ikerketa corpusari. Aurrez egin bezala, adostutako irizpideen arabera etiketatu genituen puntuak eta aldaketa interesgarria antzeman genion lortutako datu multzoari. Zuzen genbiltzan murrizte-teknika honi dagokionean, alegia, corpus eskergekin lortzen ditu daturik onenak. Gure saiakerak ordea, bestelako asmo eta ahalbideak zituen, etiketatutako corpus txiki batean oinarrituta beste oparoago bat automatikoki sailkatzea, eta lan horretan hobekuntza aipagarri bezain lagungarriak izanagatik, ez dira hain erabakigarri gertatu.

Kontuak kontu, aipagarria da algoritmo guztiekin lortu genuela *LSA* metodoarekin iragazitako 291 instantzien sailkapena hobetzea. Onenak, beste behin, *K-nn* algoritmoan oinarritutakoek eskaini zizkiguten: *F-measure* --> % 62,7.

LEKUA		**Ebaluazio neurria**	**Dimentsio-murrizketa**						
			Atributu aukeraketa				**Atributu konbinaketa**		
			Bat ere ez	**InfoGain**	**GainRatio**	**ChiSquare**	**PCA**	**LSA**	**LSA+15**
Algoritmo sailkatzaileak	**1-nn**	Performance	% 87,2624	% 90,0943	% 89,6226	% 89,6226	% 88,6792	% 85,3774	% 91,7526
		F-measure	% 84,8	% 85,8	% 84,7	% 84,7	% 85	% 84,9	% 87,8
	5-nn	Performance	% 65,566	% 83,9623	% 83,0189	% 81,6038	% 72,1698	% 86,7925	% 91,7526
		F-measure	% 84	% 84,7	% 84,7	% 84,7	% 84,7	% 87,2	% 87,8
	BN	Performance	% 89,6226	% 89,6226	% 89,6226	% 89,6226	x	% 89,6226	% 91,7526
		F-measure	% 84,7	% 84,7	% 84,7	% 84,7	x	x	x
	J48	Performance	% 89,6226	% 89,6226	% 89,6226	% 89,6226	% 78,7736	% 89,6226	% 91,7526
		F-measure	% 84,7	% 84,7	% 84,7	% 84,7	% 79,4	% 84,7	% 87,8
	NB	Performance	% 89,6226	% 90,566	% 90,566	% 90,566	% 68,8679	% 89,6226	% 91,7526
		F-measure	% 84,7	% 86,9	% 86,9	% 86,9	% 74,1	% 84,7	x
	SMO	Performance	% 89,1509	% 90,0934	% 89,1509	% 90,0943	% 88,6792	% 89,6226	% 91,7526
		F-measure	% 84,5	% 85,8	% 84,5	% 85,8	% 84,2	% 84,7	% 87,8

9. taula: 2. kategoria/klasea (***LEKUA***)

HAP

Lehenengo kategoriari zegokion taulaz geniharduenean, emaitza itxuraz apartak etiketatutako puntuen proportzioak aipatzerakoan nabari zen desorekari egozten bagenizkion, are agerikoagoa da *Lekua* gisa sailkatutako puntuen kasuan. Izan ere, 22 baitira halakotzat jo ditugunak, hala, nahiz eta sailkatzaileak baiezko batere asmatu ez, emaitza ona lortu duela dirudi. Horregatik begiratzen diogu, hain zuzen ere, *F-measure*ri.

Badira, hala ere, aipatu beharreko zenbait alderdi egindako azterketan adierazgarri direnak. Oro har, sailkatzaile guztiek antzeko emaitzak erdiesten dituzte. Inolako iragazkirik gabe egindako saiakeratan onena distantzia bakarrerako testuinguruari begiratzen dion *K-nn* algoritmoa izan da (*1-nn*), % 84,8 *F-measure* neurriaren arabera.

Dimentsio-murrizketa iragazkiak ezarri ostean, *Naive Bayes* ageri zaigu sailkatze lanetan finen. *InfoGain*, *GainRatio* eta *ChiSquare* dimentsio murriztaileak baliatuta % 90etik gorako errendimendua erakutsi du. Haatik, eta performantzia aldetik datu ez hain zapaltzailea erakutsi arren, *F-measure* datu hobea lortu dugu *LSA* eta *K-nn* erabilita (% 87,2). Alegia, guk *lekua* gisa etiketatutako puntu gehiago ezagutu ditu azken sailkatzaile horrek.

Azkenik, *LSA*ren onurak ikusita haren lana errazteko egindako saiakeran, hots, corpus zabalagoarekin egindako azterketan emaitzarik onenak lortzeko aukera izan dugu (% 87,8).

PUBLIKOA		Ebaluazio neurria	Dimentsio-murrizketa						
			Atributu aukeraketa				Atributu konbinaketa		
			Bat ere ez	InfoGain	GainRatio	ChiSquare	PCA	LSA	LSA+15
Algoritmo sailkatzaileak	1-nn	Performance	% 62,7358	% 83,9623	% 83,0189	% 81,6038	% 56,6038	% 50,4717	% 76,2887
		F-measure	% 66	% 79	% 77,2	% 74,2	% 60,8	% 55,6	% 73,6
	5-nn	Performance	% 65,566	% 83,9623	% 83,0189	% 81,6038	% 72,1698	% 55,6604	% 76,2887
		F-measure	% 68	% 79	% 77	% 74,2	% 71	% 60,1	% 74
	BN	Performance	% 80,6604	% 81,1321	% 81,6038	% 81,6038	x	% 80,6604	% 82,8179
		F-measure	% 72	% 74,6	% 74,9	% 74,9	x	x	x
	J48	Performance	% 83,4906	% 83,4906	% 84,9057	% 83,4906	% 66,5094	% 80,6604	% 82,82
		F-measure	% 79,6	% 79,6	% 80,7	% 79,6	% 67	% 72	% 75
	NB	Performance	% 80,6604	% 84,9057	% 84,434	% 84,9057	% 59,9057	% 80,6604	% 82,82
		F-measure	% 72	% 80,7	% 79,9	% 80,7	% 63,9	x	x
	SMO	Performance	% 78,3019	% 87,2642	% 87,7358	% 83,4906	% 78,3019	% 80,6604	% 82,82
		F-measure	% 75,4	% 85,2	% 85,3	% 79,9	% 72,3	% 72	% 75

10. taula: 3. kategoria/klasea (***PUBLIKOA***)

Hirugarren lema-poltsarekin lortutako datuak azaltzen dira 10. taulan. *Publikoa*ri erreferentzia egiten dieten bertso puntuak identifikatu behar zituzten aukeratu ditugun sailkatzaileek. Emaitzak ehuneko laurogeiren inguruan kokatu dira oro har.

Oraingoan adierazgarria izan da iragazkiek asko hobetu dutela haiek gabe sailkatzaileek eskaintzen zuten eraginkortasuna. Joera orokorraren kontra, baina gure kasuan ohiko emaitzak emanez, *PCA* iragazkia daukagu. Hobetu beharrean are sailkatze gaitasun okerragoa behatu dugu baliabide hau tarteko izan denean. *GainRatio* iragazkiak eta *SMO* algoritmoak osatutako bikoteak oso emaitza onak eskaini ditu (% 87,7358). Gure lanean egokientzat izan dugunak, *LSA + k-nn* (1) konbinazioak ez du klase honentzat emaitzarik onenak eskaintzerik izan, % 75eko datua jaso baitugu *F-measure* neurriarentzat. Aipatzekoa da informazio gehigarria zuen corpusarekin lortu duela azken datu hori.

Esan behar *GainRatio + SMO* sailkatzaileak lortutako balizko doitasuna ezerezean geratzen dela nahaste matrizean benetan *publikoa*ri zegozkion puntu identifikatuak 10

baino ez zirela ikusita.

a	b	<-- classified as
160	11	a = 0
31	10	b = 1

11. taula: *PUBLIKOA* klasearen *confussion matrix* edo kontingentzia-tauletariko bat.

SAIOA		Ebaluazio neurria	Dimentsio-murrizketa						
			Atributu aukeraketa				Atributu konbinaketa		
			Bat ere ez	InfoGain	GainRatio	ChiSquare	PCA	LSA	LSA+15
Algoritmo sailkatzaileak	1-nn	Performance	% 60,8491	% 64,6224	% 63,2075	% 69,3396	% 59,9057	% 66,0377	% 67,354
		F-measure	% 59,2	% 62,3	% 61,3	% 64,8	% 56,6	% 59,9	% 61,4
	5-nn	Performance	% 67,4528	% 73,1132	% 72,6415	% 73,1132	% 70,283	% 66,0377	% 67,354
		F-measure	% 58,4	% 66,2	% 63,5	% 64,5	% 58,5	% 59,9	% 61,4
	BN	Performance	% 67,4528	% 66,5094	% 68,3962	% 66,9811	x	% 69,8113	% 71,8213
		F-measure	% 58,4	% 57,9	% 56,7	% 56	x	x	x
	J48	Performance	% 68,8679	% 66,9811	% 68,8679	% 67,9245	% 57,5472	% 69,8113	% 71,8213
		F-measure	% 56,9	% 58,8	% 59,9	% 59,3	% 57,9	% 57,4	% 60
	NB	Performance	% 69,8113	% 72,1698	% 72,1698	% 716981	% 52,3585	% 69,8113	% 71,8213
		F-measure	% 57,4	% 63,2	% 62,6	% 62,3	% 54,2	x	x
	SMO	Performance	% 69,3396	% 66,5094	% 52,8302	% 66,9811	% 67,9245	% 69,8113	% 71,8213
		F-measure	% 63,5	% 60,2	% 54,1	% 59,4	% 57,9	% 57,4	% 60

12. taula: 4. kategoria/klasea (***SAIOA***)

Honakoak *saioa* klaseari erreferentzia egiten dioten puntu gisa etiketatutako instantziak sailkatzen gure algoritmoek eta iragazkiek izandako emaitzak dira. Berdintasun handia nabaritzen da algoritmo sailkatzaile guztien artean. Iragazkien bidez, hots, dimentsio-murrizketa baliatuta nahiko emaitza onak lortu dira, haiek gabe erdietsitakoak hein batean hobetuta.

ChiSquare iragazkia ezarri ondoren lortzen dira doitasun tasarik onenak eta batik bat K-nn (1) algoritmoaren bidez. Konbinazio horrekin % 64,8ko datua lortzen baita *F-measure* neurritzat.

Gainerako kasuetan bezala *LSA* iragazkiak corpus handiagoarekin eta instantzia kopuru eskergekin emaitza hobeak eskaintzen dituela nabarmen geratu da. Hain zuzen

ere, errendimendu onena eskaini digu (% 71,8213 *BN*, *J48*, *NB* eta *SMO*rekin). Haatik, etiketatuak identifikatzen baino halakoak ez direnei antzematen lortzen du hain emaitza ona, bestalde, bere *F-measure* neurriak aditzera argi ematen digunez. Nolanahi ere, sailkatzaile guztiekin lortzen dituen emaitzek izaera sendoaren berri ematen dute.

NORBERA		Ebaluazio neurria	Dimentsio-murrizketa						
			Atributu aukeraketa				Atributu konbinaketa		
			Bat ere ez	InfoGain	GainRatio	ChiSquare	PCA	LSA	LSA+15
Algoritmo sailkatzaileak	1-nn	Performance	% 69,8113	% 54,2453	% 55,6604	% 56,1321	% 43,3962	% 53,7736	% 60,4811
		F-measure	% 57,4	% 54,5	% 56	% 55,2	% 36,7	% 52,1	% 52,6
	5-nn	Performance	% 62,7358	% 62,2642	% 58,9623	% 58,4906	% 41,0377	% 51,4151	% 60,4811
		F-measure	% 58,5	% 54,7	% 43,7	% 46,6	% 23,9	% 51,3	% 52,6
	BN	Performance	% 57,0755	% 61,3208	% 57,0755	% 58,9623	x	% 58,9623	% 62,8866
		F-measure	% 56,4	% 59,6	% 51	% 56,1	x	x	x
	J48	Performance	% 59,434	% 62,2642	% 59,9057	% 61,3208	% 51,8868	% 58,9623	% 62,8866
		F-measure	% 51,3	% 57,8	% 53	% 56,4	% 51,1	% 43,7	% 48,6
	NB	Performance	% 55,1887	% 56,1321	% 57,5472	% 60,8491	% 59,434	% 58,9623	% 62,8866
		F-measure	% 52,6	% 52,2	% 44,6	% 55,3	% 58,6	x	x
	SMO	Performance	% 61,7925	% 62,2642	% 59,9057	% 61,3208	% 58,0189	% 58,9623	% 62,8866
		F-measure	% 57,1	% 60,2	% 58,2	% 58,2	% 57,6	% 43,7	% 48,6

13. taula: 5. kategoria/klasea (***NORBERA***)

Hastapenetik bereizi ditugun kategorien artean bosgarrena sailkatzeko gaitasuna egiazta daiteke taula honetan *Norbera*. Ezin esan aparteko emaitzarik erdiesten denik. Inondik ere ez, behinik behin, corpus eskergak automatikoki etiketatzeko baliatu nahi genuen tresna osatzeko asmoan lagungarri izateko lain.

Algoritmo sailkatzaileak soilik erabiliz gero distantzia bakarrera dauzkan gertueneko hitzak aintzat hartzen dituen *K-nn* (1) algoritmoa izan da arrakastatsuena zuzen sailkatutako instantzien ehunekoari begira (%69,81). Errendimenduan alde handia erakutsi arren, ez du horrenbestekoa lortu zinezko zehaztasunean *F-measure* neurrian dakusagunez (*K-nn* (5)ak hobea % 57,4 vs % 58,5).

Egindako esperimentuetan emaitzarik onenak *InfoGain* iragazkiarekin erdietsi ditugu. Hautatutako instantzia haiek sailkatzen, berriz, *sare bayestarrak (NB*) eta *SMO*

algoritmoa aritu dira egokien (*F-measure* → % 60,2).

Aipagarri iruditu zaigu *PCA* iragazki ezarri ondoren *K-nn* sailkatzaileek oso okerrera egiten zuten legez, emaitza dotoreak lortu ditugula *J48* zuhaitzekin, *sare Bayestar*rekin eta *SMO*rekin.

BETELANA		Ebaluazio neurria	Dimentsio-murrizketa						
			Atributu aukeraketa				Atributu konbinaketa		
			Bat ere ez	InfoGain	GainRatio	ChiSquare	PCA	LSA	LSA+15
Algoritmo sailkatzaileak	1-nn	Performance	% 84,434	% 87,7358	% 87,7358	% 87,7358	% 84,9057	% 72,6415	% 90,378
		F-measure	% 81,2	% 83,2	% 83,2	% 83,2	% 81,5	% 75,1	% 85,8
	5-nn	Performance	% 87,2642	% 87,2642	% 87,2642	% 87,2642	% 87,2642	% 87,2642	% 90,378
		F-measure	% 81,3	% 81,3	% 81,3	% 81,3	% 81,3	% 81,3	% 85,8
	BN	Performance	% 87,2642	% 86,7925	% 86,7925	% 86,7925	x	% 87,2642	% 90,378
		F-measure	% 81,3	% 81,1	% 81,1	% 81,1	x	x	x
	J48	Performance	% 86,3208	% 85,8491	% 85,8491	% 86,3208	% 82,0755	% 87,2642	% 90,378
		F-measure	% 80,9	% 80,6	% 80,6	% 80,9	% 81,1	% 81,3	% 85,8
	NB	Performance	% 87,2642	% 86,7925	% 86,7925	% 86,7925	% 68,3962	% 87,2642	% 90,378
		F-measure	% 81,3	% 81,1	% 81,1	% 81,1	% 72,1	x	X
	SMO	Performance	% 86,7925	% 87,2642	% 87,2642	% 87,7358	% 85,8491	% 87,2642	% 90,378
		F-measure	% 82,6	% 83,6	% 82,9	% 83,2	% 80,6	% 81,3	% 85,8

14. taula: 6. kategoria/klasea (***BETELANA***)

Azken multzo honetan gainerako taldeetan sar ez zitezkeen puntuak bildu ditugu. Saski-naski gisa ere defini genezake, gorago azaldu dugun bezala. Gutxi ziren benetan era horretara ezaugarrituak eta hori izan da, hain zuzen ere, goragoko taulan erakusten diren emaitza harrigarri onak azaltzearen arrazoia. Baiezko batere asmatu gabe ere, ehuneko laurogeitik gorako errendimendua lortzen baitute sailkatzaileek.

Hitzen artean egon daitekeen erlazio semantikoa aztertzen duen *LSA*k lortutako emaitza harrigarrian esaterako ez da egiaz betelantzat identifikatutako puntu bakar bat ere sailkatu, ikus bere nahaste matrizea:

a	b	<-- classified as
263	0	a = 0
28	0	b = 1

15. taula: *BETELANA* klasearen *confussion matrix* edo kontingentzia-tauletariko bat.

Iragazkirik ezarri gabe *SMO* algoritmoari egiaztatu diogu emaitzarik onena. *F-measure* neurriaren arabera % 82,6ko datua eskaini du.

3.9.1. Ahalegin gehigarria terminoen pisuaren arabera (Tf-idf)

Atribututzat geneuzkan hitz erro guztiei balio bera eman gabe, maiztasun eta esangura faktoreak gogoan dituen baliabide estatistiko hau gure lanean ere baliagarri izan zitekeela iruditua, iragazki gisa erabili genuen gainerako dimentsio murriztaileen osagarri eta sailkatzaileen lana hobetu asmoz. Ondoko tauletan kategoria bakoitzerako emaitza esanguratsuenak jaso ditugu.

MEZUA		Ebaluazio neurria	Dimentsio-murrizketa		
			InfoGain	GainRatio	ChiSquare
Algoritmo sailkatzaileak	1-nn	Performance	% 60,8491	% 60,8491	% 60,8491
		F-measure	% 57,7	% 57,7	% 57,7
	5-nn	Performance	% 68,3962	% 68,3962	% 68,3962
		F-measure	% 56	% 56	% 56
	BN	Performance	% 67,9245	% 67,9245	% 67,9245
		F-measure	% 55	% 55	% 55
	J48	Performance	% 67,9245	% 67,9245	% 67,9245
		F-measure	% 55	% 55	% 55
	NB	Performance	% 63,2075	% 63,2075	% 63,2075
		F-measure	% 58,6	% 58,6	% 58,6
	SMO	Performance	% 66,9811	% 66,9811	% 66,9811
		F-measure	% 59,3	% 59,3	% 59,3

16. taula: *Mezua (Tf-idf)*

Mezua etiketadun puntuak sailkatzeko azken ahalegin honetan begien bistakoa da

dimentsio-murrizketarako iragazkiek (aurresan zitekeen bezala, bestalde) ez dutela inolako eraginik sailkatzaileen eraginean. Termino bakoitzaren pisu ponderatua aintzat hartuz gero, aurrez ere datuak bahetu ditugunez, gure saiakeran ez zegoen hori erabilitako baliabideen bidez gehiago fintzerik.

Beste behin ere, errendimendu portzentajeak eta *F-measure* neurriak ondorio ezberdinetara garamatzatela ikusten da azterketa honetan. *K-nn* (5) algoritmoarekin errendimendu portzentajerik onena lortzen bada, *SMO*rekin erdietsi dugu emaitzarik baliagarriena asmatze tasa bateratuari so (% 59,3).

Hurrengo tauletan errendimendu datuak (*performance*) ezkutatu ditugu iragazkien emaitza errepikatuekin batera, izan ere, azaldu berri dugun arrazoien kariaz berdintzen diren edo baliagarri ez den informazioarekin lana nahasi eta astundu baitaiteke.

LEKUA		**Ebaluazio neurria**	**Dimentsio-murrizketa**
			InfoGain/GainRatio/ChiSquare
Algoritmo sailkatzaileak	**1-nn**	F-measure	% 84,7
	5-nn	F-measure	% 84,7
	BN	F-measure	% 84,7
	J48	F-measure	% 84,7
	NB	F-measure	% 84,2
	SMO	F-measure	% 85,6

17. taula: *Lekua (Tf-idf)*

SMO algoritmoarekin lortzen dira emaitzarik onenak *lekua* identifikatze lanetan. Ez dago sailkatzaileen artean alde handirik emaitzei dagokienean, gorago esan bezala nahikoa bateratzen baitira guztien errendimenduak.

PUBLIKOA		Ebaluazio neurria	Dimentsio-murrizketa: InfoGain/GainRatio/ChiSquare
Algoritmo sailkatzaileak	1-nn	Performance	% 74,0566
		F-measure	% 75,5
	5-nn	Performance	% 81,6038
		F-measure	% 74,2
	BN	Performance	% 80,6604
		F-measure	% 72
	J48	Performance	% 80,6604
		F-measure	% 72
	NB	Performance	% 84,9057
		F-measure	% 83,2
	SMO	Performance	% 84,9057
		F-measure	% 81,9

SAIOA		Ebaluazio neurria	Dimentsio-murrizketa: InfoGain/GainRatio/ChiSquare
Algoritmo sailkatzaileak	1-nn	Performance	% 48,5849
		F-measure	% 50,6
	5-nn	Performance	69,8113
		F-measure	% 57,4
	BN	Performance	% 69,8113
		F-measure	% 57,4
	J48	Performance	% 69,8113
		F-measure	% 57,4
	NB	Performance	% 47,1698
		F-measure	% 49
	SMO	Performance	% 67,9245
		F-measure	% 56,5

18. eta 19. taulak: *Publikoa* eta *Saioa (Tf-idf)*

Ezberdintasun nabariagoak agertu zaizkigu *Publikoa* eta *Saioa* kategoriak identifikatze ahaleginetan. Lehenengoari dagokionean *Naive Bayes* algoritmoak eskaintzen du doitasunik handiena. *Saioa* sailkatzerakoan, aldiz, *K-nn* (5), *Sare Bayestarrak (BN)* eta *J48* zuhaitzek lortu dituzte daturik hoberentsuenak.

NORBERA		Ebaluazio neurria	Dimentsio-murrizketa: InfoGain/Gain Ratio/ChiSquare
Algoritmo sailkatzaileak	1-nn	F-measure	% 50,2
	5-nn	F-measure	% 44,1
	BN	F-measure	% 43,7
	J48	F-measure	% 43,5
	NB	F-measure	% 52,4
	SMO	F-measure	% 49,9

BETELANA		Ebaluazio neurria	Dimentsio-murrizketa: InfoGain/Gain Ratio/ChiSquare
Algoritmo sailkatzaileak	1-nn	F-measure	% 83,2
	5-nn	F-measure	% 81,3
	BN	F-measure	% 81,3
	J48	F-measure	% 81,3
	NB	F-measure	% 78,7
	SMO	F-measure	% 82,2

20. eta 21. taulak: *Norbera* eta *Betelana (Tf-idf)*

HAP

Azkenik, *Norberari* erreferentzia egiten dioten puntuak ezagutzeko zailtasun nabarmenak izan ditugula erakusten duen taulan, emaitzarik onena *Naive Bayes* algoritmoak eskaini digu. *Betelana* zehazteko ahaleginetan berriz, *K-nn*(1) izan dugu aukerarik onena.

Begirada zabala eginda, *TF_IDF* bidez berariaz hautatutako atributuekin egindako ahaleginak ez ditu hobetu aurreko emaitzak, aitzitik, zenbaitetan datu okerragoak ere ematen ditu. Oro har, zenbatekoak bateratu edo orekatu egin dira eta terminoen ponderazioak gainerako iragazkien eragina erabat murriztu du. Beraz, hurrengo urratsa emateko aukera berririk ez digu eskaini metodo estatistiko honen aplikazioak.

3.9.2. Multi-sailkatzaileekin lortutako emaitzak

Orain arteko datuak ikusita sailkatzaile batzuek besteek baino emaitza hobeak erdiesten dituztela egiaztatu dugu eta berdin esan daiteke iragazki gisa erabili ditugun dimentsio murriztaileekin. Jakina da, bestalde, badagoela sailkatzailerik onenak multzokatzerik multi-sailkatzaile bat egiteko. Ereduak elkarri atxiki dakizkioke. Datu multzo beraren gainean, hainbat sailkatzaile jarri eta tarteko hipotesiak lortu ostean, elkarren konbinazioa lortzea da bidea. Hari horri jarraiki, azken xedea sailkatzaile egonkorragoa, zehatzagoa aurkitzea litzateke.

Dagoeneko 3.5 atalean azaldu dugun bezala, *Vote*, *Adaboost* eta *Bagging* tekniken artean azken hau proposatu zaigu geniharduen arloan egokien izan zitekeen baliabidetzat. Gehiengoaren bozka zein batez bestekoa hartzen ditu irizpidetzat, unean unean egokiena dena. Hainbat sailkatzailek egindako irizpenen batez bestekoa egiten du hasierako datu multzotik eratorritako hainbat lagin behin eta berriz aztertu ondoren.

Metodo honekin bariantza handiko hainbat sailkatzaile konbinatu eta izan dezaketen errorea murrizten da (esaterako sailkapen-zuhaitzetan gertatzen dena). Hainbat lanetan ezagutu ditugu hobekuntzak [Breiman, 1996] eta erabilera erraza izanik zein gure ikasketa metodoari egokitzen zitzaiola iritzita, saiatzea erabaki genuen.

Bagging metodoarekin lortutako emaitza sorta bat ikus daiteke 22. taulan. Klase edo kategoria bakoitza aztertu eta haietan emaitzarik onenak lortzen zituen sailkatzailea,

hots, aurreko esperimentuetan *F-measure* balio altuena zuena aukeratu dugu. Goiko lerroan klasearen zenbakia adierazi dugu (*mezua, lekua, publikoa, saioa, norbera, betelana*) lan guztian jarraitutako hurrenkeran. Algoritmo mota eta dimentsio-murrizketa edo atributuak aukeratzeko metodoa *Bagging* izeneko lerroan ikus daiteke.

Bagging	1 (*Mezua*)		2 (*Lekua*)		3 (*Publikoa*)		4 (*Saioa*)		5 (*Norbera*)		6 (*Betelana*)	
	IB1_None_80	IB1_None_100	SMO_InfoGain_80	SMO_InfoGain_100	SMO_ChiSquare_80	SMO_ChiSquare_100	IB5_None_80	IB5_None_100	SMO_InfoGain_80	SMO_InfoGain_100	IB1_GainRatio_80	IB1_GainRatio_100
Performance	% 57,075	% 54,245	% 89,623	% 89,623	% 82,075	% 82,075	% 69,34	% 67,924	% 58,906	% 58,019	% 87,736	% 87,736
F-measure (0)	% 68,7	% 63,9	% 94,5	% 94,5	% 89,9	% 89,8	% 81,8	% 80,8	% 71,4	% 71,7	% 93,4	% 93,4
F-measure (1)	% 31,6	% 37,4	0	0	% 17,4	% 26,9	% 3	% 2,9	% 24,1	% 18,3	% 7,1	% 13,3

22. taula: *Bagging*-aren emaitza

Bitarteko honek algoritmoa aldatu gabe, datu-multzoak edo behaketarako erabili asmo dugun lagina hainbat aldiz biderkatzeko modua eskaintzen du. Horixe da, hain zuzen ere *Bagging* lerroko izendapenetan azkeneko zenbakiak adierazten duena. Esaterako, *IB1_None_80*-eko *80* horrek laginaren edo *lemma-poltsa*ren permutazio edo biderkatze neurria ematen du. Haiekin egindako aldaketak behatuta, errendimenduak ez du hobera egin lagintze anitzagoa aintzat hartuta ere. Alabaina, egiaz markatu beharreko ezaugarriak asmatzeko aukera gehiago eskaintzen duela dirudi, esan nahi baita, egiaz etiketatu eta ezagutu beharreko bertso puntuak sailkatu edo antzematerakoan aztiago diharduela metodo honek, hein batean bada ere. Ikus *mezua*, *publikoa* eta *betelana* klaseetako *F-measure*(1) adibideak. Kontrakoa ikusten da, nolanahi ere, *saioa* eta *norbera* etiketak antzematerakoan.

3.9.3. Emaitzarik nabarmenenak

Ahalegin bakoitzarekin lortutako datuen berri eman dugu. Saiakera ugari izan dira eta tauletan bildutako datuak ere askotarikoak. Hori dela eta, egokia iruditu zaigu laburpen

gisa eta ondorioen atalaren aurretik klase bakoitzarentzako daturik onenen taula bat eskaintzea.

Eskuineko zutabean *LSA* + 15 lagin oparoagoa eta dimentsio-murrizketarako teknika ezagunarekin inoiz erdietsi dugun emaitza hobeagoa jarri dugu saiakeraren osagarri gisa.

	F-measure	Algoritmoa eta dimentsio-murrizketa metodoa	*LSA*+15
Mezua	% 62,3	*K-nn* (1) Dimentsio-murrizketarik gabe	% 62,7
Lekua	% 86,9	*Naive Bayes* *InfoGain*	*% 87,8*
Publikoa	% 85,3	*SMO* *GainRatio*	
Saioa	% 64,8	*K-nn (1)* *ChiSquare*	
Norbera	% 60,2	*SMO* *InfoGain*	
Betelana	% 83,6	*SMO* *InfoGain*	*% 85,8*

23. taula: Emaitzarik nabarmenenak

4. Erregelekin egindako saioak

Emaitzek erakutsi dute hobetzeko tarte handia dagoela. Izan ere, lortutako emaitzekin ez baitago etiketatu gabeko corpus zabalak sailkatzerik eta finean, bertsolaritzan eta zehazki agur-bertsoetan egitura jakinak, estandarrak antzemateko lanari ekiterik. Gauzak horrela, hasierako eztabaidetan proposatutako beste bide bati berriz heltzea deliberatu genuen. XX. mendeko laurogeita hamarreko hamarkadara arte sailkatzaile automatikoak

egituratzeko eskuarki baliatzen ziren sistema adituen bidean, ahalegin berri bat egin genuen: Balizko sailkatzaileak geronek definitutako erregelen bidez ezagutu behar zituen ezarritako sei kategoriak.

Esan genezake ikasketa automatiko gainbegiratuaren barrenean hurbilketa bitarra egin dugula, erregela induktiboak eta instantzietan oinarritutako ikasketa sistemak uztartuta. Azken xedea jakina da, batez besteko doitasuna areagotzea zeina F-measure neurriaren bidez egiaztatuko dugun.

4.1. Erabilitako baliabideak

Aipatu erregelak osatzeko aztertutako bertso-puntu edo instantzietan antzeman ditugun ezaugarriak harrapatu nahi genituen. Hainbat aldiz begiztatutako corpus horretan arrantzan aritu behar genuen. Sarea prestatzeko Euskararen Datu-Base Lexikala (EDBL) baliatzea erabaki genuen. Bertako ezaugarri eta etiketak izan ditugu gogoan klaseak biltzeko sarea bilbatzeko garaian. Aipatu etiketak lagintzat hartutako corpusean esleitzeko, berriz, EUSTAGGER tresna izan dugu lagungarri.

EDBL (Euskararen Datu-Base Lexikala) euskarazko testuak prozesatzeko lanetan erabiltzen den erabilera anitzeko datu-base bat da. Egun 80.000 sarreratik gora biltzen ditu eta hizkuntza naturalaren azterketan diharduten hainbat ahaleginetan oinarrizkoa izan da. Hizkuntz teknologien alorreko hainbat tresnentzako eskaintzen du informazio lexikoa: analisi morfologikoa, ortografia egiaztatze nahiz zuzenketa, lematizatze eta etiketatzea, analisi sintaktikoa eta abar. Hizkuntzalaritzako zenbait formalismorekiko neutral izateko diseinatu da eta informazio mota berriak aintzat hartzeko bezain malgua.

EUSTAGGER estaldura zabala eskaintzen duen Euskararen analizatzaile morfologikoa eta kategoria ezarlea da. Analizatzailea bi mailako formalismoan oinarritzen da eta modu inkrementalean hiru modulutan egituratu dute: analizatzaile estandarra, aldaera linguistikoen analizatzailea eta hiztegirik gabeko analizatzailea, zeinak lexikoian bildu gabeko hitzak ezagutu ditzakeen. Desanbiguatze metodo gisa murriztapen-gramatika eta Markoven eredu ezkutuan oinarritutako etiketatzailea darabiltza. Finean, zehaztasun handiko tresna dela esan behar da (% 3,5-eko errore tasa).

4.2. Erregelak

Jarraian, gorago aipatu dugun sarea josteko erabili ditugun hari mutur eta korapiloak eskaintzen ditugu. Kategoria bakoitzarentzat erregela edo bilaketarako etiketa zerrenda bat adostu dugu behaketak hala iradokita.

a) Mezua: Kategoria edo klase hau identifikatzeak berezko dituen zailtasunak ezein mezuren mamia jendarteratzerakoan bertsolariak behar duen askatasunak berak dakartzanak dira. Hori dela eta bertsoen egituran maiz proposatu den, erabilera ohikoak aditzera ematen duen eta aditu zein irakasleek proposatzen duten ildoari lotu gatzaizkio. Bertsoetako azken puntua izaten da eskuarki haren mamia, esanahia, eztena nahiz mezua biltzen duena, beraz, azken puntu oro *mezu* etiketapean ezartzen duen erregela egin dugu.

b) *Lekua*: Kokagune bati erreferentzia egingo lioketen puntuak atzemateko EDBLn jasotzen diren ezaugarri morfologiko eta etiketen artean leku-izen bereziak bururatu zitzaizkigun lehendabizi. Toki horiek agertzeak baliteke ezinbestean ez ekartzea gune horietaz ari den puntua ezagutzeko modua; haatik, agerpen horrek zantzu nahiko esanguratsua eskaintzen du. Bestalde, lekuzko postposizioak ere baliatzeko modukoak zirela uste izan genuen saiakeraren hastapenetan. Hona bilaketan baliatu genituen etiketak: LIB, ALA, ABL, INE. Leku-izen bereziekin batera leku denborazko *adlatibo (-ra)*, *ablatibo(-tik)* eta *inesibo (-n)* postposizioen laburdurak dira, hurrenez hurren.

Azkenean, eta emaitzak aztertu ondoren, estalduraz ahalmen handia zuen erregela hau leku-izen berezien (LIB) bilaketara mugatu genuen doitasunaren mesedetan.

c) *Publikoa*: Aurrean duen jendarteari zuzentzerakoan bertsolariak bigarren pertsona erabili ohi du, nahikoa gertutasun adierazteko eta beti ere begirunez. Hori dela eta EDBLn zehaztutako etiketen artean pertsona horietako izenordeak nahiz adizkietako morfema jakinok ezaugarritzeko erabili direnak hartu ditugu aintzat: bigarren pertsona singular nahiz pluraleko morfemak, alegia, *Nork, Nori* eta *Nor* kasuetakoak (NK_ZUK, NK_ZUEK, NI_ZURI, NI_ZUEI, NR_ZU, NR_ZUEK) eta 2. pertsonako izenordain arruntak zein indartuak (IOR PERARR ZU, IOR PERARR ZUEK, IOR PERIND ZU, IOR PERIND ZUEK).

HAP

Gainera, klase bakoitzerako berariazko hiztegitxo bat osatzea erabaki genuen, ahaleginak ahalegin oraindik ere gure galbaheak jasotzen ez zituen puntu ale ustez garrantzitsu edo esanguratsuak biltze aldera. Adituen gomendioei eta lanean metatu dugun eskarmentuari jarraiki, adostutako hiztegi horietan termino bakan batzuk ezarri ditugu. Hona *publikoa* kategoriarentzat hautatutakoak: *entzule, txalo, bertsozale*.

d) Saioa: Kategoria honetan ekintzako parte-hartzaileak eta burutu beharreko atazak hartzen dira kontuan. Hortaz, behaketa ugariren ostean izen bereziek garrantzi handiko egitekoa osatzen dutela ohartuta, horiek hartu ditugu erregelako irizpide nagusi gisa (EDBLko IZB etiketa). Halaber, hirugarren pertsonako izenordain batzuk bildu beharra ageriko suertatu zaigu saioari erreferentzia egiterakoan maiz baliatzen direlako, esaterako, *beraiek* (IOR PERARR HAIEK).

Era honetako ideiak jasotzen dituzten puntuak harrapatzeko osatu dugun hiztegian berdin, agerraldiotan esanahi berezia duten elementuak bildu ditugu termino ezaugarritzaile gisa. Saioaren egiturari eta bertako partaideei lotutako hiztegi hau baliatu dugu: *goiz, arratsalde, gai-jartzaile, epaitu, entrenamendu, epaile, txapelketa, bertso, ariketa, saio, ohore, gaur, kantu, kanta, txapela.*

e) Norbera: Erreferentzia hurbilena azaltzen duten puntuotan lehenengo pertsonako izenordainak eta morfemak bilatu ditugu EDBLko etiketak erabilita. Subjuntiboa (norbere nahiak adierazteko erabiltzen baita), moduzko adierazpideak eta adizki jakinak ere hartu nahi izan ditugu kategoria hau definitzeko oinarrizko elementu gisa.

Jarraian erregelan erabilitako etiketa sorta zerrendatu dugu: Adizkietan antzematen diren lehen pertsonako morfemak (NK_NIK, NK_GUK, NI_NIRI, NI_GURI, NR_NI, NR_GU), subjuntiboko formak (A3), lehen pertsonako izenordainak (IOR PERARR NI, IOR PERARR GU), *behar* (L-A-IZE-ARR-1100, L-A-IZE-ARR-1307) edo *nahi* + aditz laguntzailea (ADI HAUL_EDBL). Lokailu kontzesiboekin egindako saiakera aurreragoko lanetarako utzi dugu (LOT LOK KONT (*nahiz_eta*)).

f) Betelana: Hauek identifikatzeko *mezua* kategoriakoekin erabakitako bide berari heldu diogu, hots, erregela baino irizpide bat ezarri dugu. Hain zuzen ere, gainerako kategorietan sartzen ez diren puntu guztiak saski-naski litzatekeen multzo honetakotzat jotzea deliberatu dugu. Beraz, ez da erregelarik klase honentzat eta

besteetan biltzerik ez direnek edo erregelak liratekeen sareetan jasotzen ez direnek osatuko dute bertso-puntu poltsa hau.

4.3. Erregela bidezko ahaleginaren emaitza

Orain arte eskainitako taula eskergekin berriz ere informazio olde nahasgarririk ez genuke eskaini nahi. Horrela, lortutako emaitza guztien artean bildutako daturik onenak soilik jasotzen dira 24. taulan.

Ikasketa automatikoan erabilitako algoritmoak aurreko atazetan baliatu ditugunak beraiek izan dira. Datuak errendimenduari eta *F-measure* neurriari dagokionak dira eta erregelak sortu ditugun kategorientzat lortutako emaitzak eskaini ditugu. *Betelana* eta *Mezua* kategorien erregelak irizpide edo erabaki bati jarraiki hartu ditugu ez arau linguistikoei begira. Erabaki hori ez dugu noski ausaz hartu, ikerketa aldian egindako behaketetan ondorioztatutako zioen kariaz baizik (ikus 4.2 ataleko *mezua* eta *betelana* azpiatalak). Bat-bateko bertsoetan arrazoibiderik indartsuena eta landuena bukaerarako gorde ohi du bertsolariak [Garzia, Sarasua, Egaña, 2001]. Hain ekitaldi laburrean, berrogeiren bat segundokoa, osatu beharreko ibilbidea zehazteko astia izango du igorleak eta entzuleen artean ere amaiera indartsuak arrakastarako bidea arinduko dio. Beraz, bukaerako arrazoi hori *mezu* gisa etiketatzeak badu adituen zein arituen eta tradizioaren nahiz komunikazio arauen babesa.

Erregelekin uztartuta algoritmo sailkatzaileek lortutako emaitzarik onenak	**Performance**	**F-measure**
Mezua	% 85,5172	% 85,1
Lekua	% 97,5862	% 97,5
Publiko	% 89,3103	% 88,5
Saioa	% 74,1379	% 72,4
Norbera	% 80,6897	% 80,7
Betelana	% 87,931	% 82,3

24. taula: Erregelen emaitza

J48, NB, SMO eta algoritmo sailkatzaile gehienek emaitza bera itzultzen dute klasea erabakitzeko atributu bakarra erabiltzen dutelako.

Batik bat *lekua*ri dagozkion emaitzak onak izan dira, alabaina arrazoia nahikoa begien bistakoa da; izan ere, egindako azterketa morfologikoan oinarritutako erregelek gidatzen baitute etiketatzea. Erregela bidezko programa batekin, esan nahi baita, ikasketa automatikoko algoritmorik erabili gabe emaitza berbera lortu dugu. Ez digute bada *ikasketa automati*ko (IA) algoritmoek nahi genuen beste lagundu.

4.4. Azterketa morfologikoan oinarritutako etiketatzea

Hona sortu dugun tresnarik onenak ematen duen emaitza. Honela etiketatzen du eranskinetan jaso dugun agur-bertsoak etiketatzeko programak edo baliabideak. Batekoek klase horretako puntutzat jotzen dela adierazten du, zeroek berriz, ez dela proposatutako klasekoa.

A. Otamendi Doinua: Betroiarena	*Klaseak*					
	1 Mezua	*2 Lekua*	*3 Publikoa*	*4 Saioa*	*5 Norbera*	*6 Betelana*
Txapelketa hasi da eta gabon Leidor	0	1	0	1	0	0
gaur atzerako dira batzuk derrigor	0	0	0	0	0	1
baina ahaleginen bat ez ote dugu zor	0	0	0	0	1	0
Xebastian gaur egun ez dugu edonor	0	0	0	1	1	0
Unai fuerte dator	0	0	0	1	0	0
Legarreta gogor	0	1	0	0	0	0
Bixente egoskor	0	0	0	1	0	0
prestatuta Igor	0	0	0	1	0	0
baina ni hauek danak haustera nator	1	0	0	0	1	0

25. taula: Automatikoki etiketatutako agur-bertsoa

5. Ondorioak

5.1. Emaitzen interpretazioa

Saiakera honetan bat-bateko agur-bertsoen analisi automatikoarentzako oinarrien bila aritu gara.

Azterketa lanean agur-bertsoetako ezaugarri diskurtsibo orokorrenak proposatzen dira, orobat, aipatu ezaugarriek balizko egiturak automatikoki sailkatzerakoan osatu bide diren atazetan zenbaterainoko egokitasuna erakusten duten azalduz. Lehenetsi dugun helburua bertso-puntuak zuzen sailkatzeko ezaugarriak zedarritzea izan da; gainera, honi jarraitzen zaion jomugatzat izan dugu bertsoen egitura narratiboa analizatzea. Bigarrenik, agur-bertsoak kategorizatzeko hainbat metodoren baliagarritasuna egiaztatzen ere ahalegindu gara.

Xede hauen xerka, bat-bateko agur-bertsoetan antzeman ditugun hainbat ezaugarri bildu ditugu, zeintzuek sei multzotan kategorizatuta zehaztu eta murriztu ditugun: *Mezua*, *Lekua*, *Publikoa*, *Saioa*, *Norbera* eta *Betelana*. Ondoren, ikasketa automatikoko sei hurbilketa modu ezarri ditugu dimentsioa murrizteko teknikak ikasketa-algoritmoekin uztartuta. Hurbilketa bakoitza gorago aipatu kategoriari zegokiona izan da.

Aurreneko esperimentuetan lortutako emaitzak ez ziren azken xedea erdiesteko nahikoa egokiak. Asmatze portzentajea % 65etik % 89ra arteko tartean izanagatik, benetan axola zitzaigunik ez baitute egiterik, alegia, klase batekoak direla asmatzeko gaitasuna txikia da: NaiveBayes algoritmoaren kasuan ia batere ez du asmatzen, J48 zuhaitzek gutxi batzuk eta Knn-en oinarritutakoek zerbait gehiago (kasurik onenean 32 instantzia, txarrenean batere ez).

Datuon karietara, emaitzon arrazoien bila hasi eta atributuen ugaritasunean aurkitu dugu findu beharra duen lehen esparrua. Asmo horrekin maiztasun handieneko hitzen zerrenda pare bat jaso (hitz lauzko nahiz bertsoetako testuetan bilatuta) eta handik gure lana eragotzi gabe kentzekotzat jotzen genituen hitzen zerrenda osatu dugu, hemen *lasto-hitz* gisa bataiatuak (*StopWords*). Gainera, gerora kaltegarri gerta zitezkeen

estandarretik aldendutako hitzak egokitzea ere erabaki dugu, bikoizketak ahalik eta gehien saiheste aldera (ikus 3.2 aurre-prozesatze lana).

Aldaketa guzti hauekin *.arff* fitxategi berriak[14] osatu ditugu eta WEKAk atributuak murrizteko eskaintzen dituen tresnak baliatuta ezaugarri horiek ehunetik behera jaisten ahalegindu gara, gero emaitza hobeak erdiesteko itxaropenari eutsita.

Lortutako emaitzen karietara ez dago algoritmo bakar bat aipatzerik kategoria guztiak sailkatzeko egokiera berezirik erakusten duena. Haatik, egia da, *K-nn* algoritmoak hurbileneko distantziara zituen datuak aintzat hartuta, hots, 1-eko distantziara emaitza onak eskaini dituela *lekua*, *publikoa* eta *saioa* gisa etiketatutako instantziak sailkatzen (ikus 9., 10. eta 12. taulak). Bada ordea bigarren kategoria multzoa (*lekua*) sailkatzerakoan datu apartak eskaini dituen beste iragazkirik (*Naive Bayes*) edota hirugarrenean (*SMO*) eta laugarrenean (*BN*) antzeko arrakasta lortu dutenak ere.

Iragazki murriztaileak ezarri ostean jasotako emaitzak interpretatzerakoan, dagoeneko adierazia dugu *LSA* teknikak nahikoa sendotasun eskaini duela. Berau izan dugu ia kategoria guztietan emaitzak hobetzeko tresna baliagarri. Alabaina, hemen ere ez dago erabateko fidagarritasuna aitortzerik. Izan ere, *GainRatio*rekin asmatze portzentaje bikaina erdietsi da esaterako publikoa klasea sailkatzerakoan; *norbera* gisa etiketatutako puntuak sailkatzeko *InfoGain*ek itxura ona erakutsi du; eta beste adibide bat ematearren *ChiSquare* ere egoki moldatu da *saioa*ri erreferentzia egiten zioten puntuak bilatzen (ikus emaitzarik nabarmenenak, 23. taula).

Iragazkien izaera dela-eta, hobekuntzarako berariazko bideak ditu murrizte-teknika bakoitzak. *LSA*renak lantzen ahalegindu gara gu. Etiketatutako corpusa zenbat eta oparoagoa, orduan eta eraginkorragoa da aipatu teknika, halaxe frogatu ahal izan dugu jasotako datuetan eta garbi islatzen da lan honetan eskainitako tauletan.

5.1.1. Saiakera osagarriak

Atributuen pisua neurtzen lagungarri den TF_IDF teknikaren bidez berariaz hautatutako atributuekin egindako ahaleginak ez ditu hobetu aurreko emaitzak, aitzitik, zenbaitetan datu okerragoak ere ematen ditu. Elementu lexikoen pisua aintzat hartu izanak gerora

14 Lema-poltsak (*bag-of-lemma*) osatzeko erabili dugun fitxategi-formatua da .arff

baliatu ditugun iragazkien eragina indargabetu edo murriztu dute. Ez du gure esperimentuetan urrats berririk emateko modurik ekarri metodo honekin egindako saiakerak.

Teknika murriztaileak eta ikasketa-algoritmoak uztartzen dituen *Bagging* metodoa ere baliatu dugu, ahalegin berrien sailean. Datu-laginekin lan egiten duen metodo honekin emaitza interesgarriak lortu ditugu lagintzea biderkatuta. Saiakera bakoitzean algoritmo bakarra erabiltzen da metodo honetan eta klase bakoitzean daturik onenak lortutakoekin saiatu ginen. Benetan markatu ditugun ezaugarriak asmatzeko erraztasun handiagoa eskaintzen duela dirudi. Hobekuntza arina izanik ere, ez deritzogu erabakigarria lehenetsitako helburua (agur-bertsoetako egitura diskurtsiboa) lortze bidean (ikus 22. taula). Izan ere, ez du inoiz hobetzen bestelako saiakeratan eskuratutako datu sortarik onena (emaitzarik onenak biltzen dituen 23. taulan ez da b*agging*ekorik).

Azken ahalegina erregela bidezko saiakerarekin egin dugu. Jarraitutako prozedura dela-eta, algoritmo sailkatzaileek eskuarki emaitza parekoak itzuli dizkigute. Klasea erabakitzeko ezaugarri bakarra erabiltzen dutenez, hots, erregela bera, ondorioa ulergarria da. *Lekua* kategorizatzeko trebezia aparta erakutsi badu ere, osatu dugun azterketa morfologikoan oinarritutako erregelek etiketatzea gidatzeari egotzi behar diogu horren errua. Erregela bidezko programa batek, alegia, ikasketa automatikoko algoritmorik erabili gabe emaitza berbera lortu dugu.

5.2. Aurrera begira egin daitezkeenak

Erregelatan oinarritutako kodetze automatikoak zein ikasketa automatikoa darabilenak, biek eskaintzen dute datuak kodetzeko kalitatezko aukera, are bildutako testuen kalitate txarra kontuan izanik ere.

Oro har, erregelatan oinarritutako hurbilketak emaitza hobeak eskaini ditu ikasketa automatikoa baliatuz egindakoak baino, bereziki aztergai ziren instantziak gutxi izan direlako. Bistakoa da, jakina, adituak kalitate handiko erregelak proposa ditzakeela instantzia kopuru apurra izanik ere. Gainera, erabaki onak har ditzake kodetze lanerako ezaugarririk egokienak aukeratzeko orduan.

Lortutako emaitzak aintzat hartuta, hainbat bide jarrai genitzake etorkizunean. Aurrena, ezaugarri semantiko hobeen bila ahalegindu gaitezke, eta orobat, ikasketa

automatikoko algoritmo ugarien artean egokienak aukeratzen saiatu. Honako lana lehenbiziko urratsa da norantza horretan. Ikasketa automatikotik ateratakoa giza adituak zuzentzeko mekanismorik ere sor liteke eta erregelak fintzeko baliatu. Instantzia gehiago beharko ditugu eta hala automatikoki etiketatzeko lanak zinez hobetuko direlakoan gaude.

Instantziak ugaritzearen ildotik, bertsolaritzan (ikerlari, epaile, inprobisatzaile) diharduten adituengandik jaso ditugun iradokizunak gogoan hartu nahi ditugu. Jaialdietako bertso sortekin aritzea, koplatako puntuekin, hileta bertsoak, egitura nabariagoa duten bestelako jarduerak, adierazle linguistikoak aintzat hartzea (aditzondoak, formulak, izenordainak, aditz aspektua) edota bertsolari bikote bat guk aurrez erabakitako kategoria edo ezaugarriak errepikatuko dituzten bertsoak sortzen jartzea, besteak beste.

Bagging-arekin egin bezala, *Vote* edo antzerako beste multi-sailkatzaileekin saia gaitezke emaitza hobeak lortzen.

Finean, corpus etiketatu zabalagoak presta genitzake, sailkatzaileei aukera gehiago eskaintzeko eta finean, diskurtso egiturak antzematen ahalegintzeko.

6. Bibliografia

[Alegria et al., 1996] Alegria, I., Artola, X., Sarasola, K., and Urkia, M. (1996). Automatic morphological analysis of Basque. Literary and Linguistic Computing, 11(4):193{203.

[Arrieta, 2010] Arrieta, B. (2010). Azaleko sintaxiaren tratamenduaikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean. EHU Lengoaia eta Sistema Informatikoak saila. Doktore tesia.

[Austin, 1975] Austin, J. L. (1975). How to do things with words, volume 88. Harvard University Press.

[Black, 1991] Black E., Abney S., Flickenger D., Gdaniec C., Grisham R., Harrison P., Hindle D., Ingria R., Jelinek F., Klavans J., Liberman M., Marcus M., Roukos S., Santorini B., eta Strzalkowski T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. *Proceedings of DARPA Workshop on Speech and Natural Language.*

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. In *Machine learning*. 123-140 or.

[Cardoso-Cachopo and Oliveira, 2003] Cardoso-Cachopo, A. and Oliveira, A.(2003). An empirical comparison of text categorization methods. In String Processing and Information Retrieval, pages 183{196. Springer.

[Dasarathy, 1991] Dasarathy, B. V. (1991). Nearest neighbor (fNNg) norms:fNNg pattern classication techniques.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391{407.

[Escandell, 1996] Escandell, M.ª V. (1996). *Introducción a la pragmática. Barcelona: Ariel.*

[Euskaltzaindia, 1990] Euskaltzaindia (1990). *Euskal gramatika: lehen urratsak-III (lokailuak)*. Euskaltzaindia, Bilbo.

[Ezeiza et al., 1998] Ezeiza, N., Alegria, I., Arriola, J. M., Urizar, R., and Aduriz, I. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics, 1. liburukia*, 380-384 or. Association for Computational Linguistics.

[Foley, 2004] Foley, John. (2004). Ahozko tradizio konparatiboak in *Inprobisazioa munduan : kulturarteko topaketak 2003, txostenak.* 15-45 or. Donostia. Euskal Herriko Bertsolari Elkartea.

[Forman, 2003] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classication. The Journal of Machine Learning Research, 3:1289-1305.

[Goenaga, 1980] Goenaga, Patxi. (1980). *Gramatika bideetan.* Erein. Donostia

[Garzia, 2008] Garzia, J. (2008). *Jendaurrean hizlari*. Alberdania.

[Garzia, Sarasua, Egaña, 2001] Garzia, J., Sarasua, J., Egaña, A. (2001). *Bat-bateko bertsolaritza: gakoak eta azterbidea.* Donostia. Bertsozale elkartea.

[Gumperz, 1982] Gumperz, J. J. (1982). *Discourse strategies: Studies in interactional sociolinguistics.* Cambridge University, Cambridge.

[Hainbat egile, 2004] Hainbat egile. (2004). *Ahozko inprobisazioa munduan topaketak.* Donostia. Euskal Herriko Bertsozale Elkartea.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). *The WEKA data mining software: an update*. ACM SIGKDD Explorations Newsletter, 11(1):10{18.

[Hymes, 1972] Hymes, D. (1972) Towards ethnographies of communication: the analysis of communicative events. In Giglioli, P (ed.). *Language and Social Context.* Harmondsworth: Penguin books; 21-33 or.

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1):177{196.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98, pages 137{142.

[Kim et al., 2002] Kim, S.-B., Rim, H.-C., Yook, D., and Lim, H.-S. (2002). Efective methods for improving naive bayes text classiers. PRICAI 2002: Trends in Articial Intelligence, pages 479{484.

[Leopold and Kindermann, 2002] Leopold, E. and Kindermann, J. (2002). Text categorization with support vector machines. how to represent texts in input space? Machine Learning, 46(1):423{444.

[Márquez, 2002] Marquez L. (2002). Aprendizaje automatico y procesamiento del lenguaje natural. *Tratamiento del lenguaje natural*, 207 or..

[Minsky, 1961] Minsky, M. (1961). Steps toward articial intelligence. Proceedings of the IRE, 49(1):8{30.

[Pimienta and Trapero, 2001] Pimienta, A. and Trapero, M. (2001). Teoría de la improvisación: primeras páginas para el estudio del repentismo. Ediciones Union.

[Prospekzio soziologikoen kabinetea, 2007] Eusko Jaurlaritzako Prospekzio Soziologikoen Kabinetea, 2007ko abendua. Euskal Kultura. http://www.euskadi.net/contenidos/informe_estudio/o_08euskal_kultura/eu_euskul/adjuntos/07euskalkultura.pdf

[Quinlan, 1993] Quinlan, J. R. (1993). C4. 5: programs for machine learning, volume 1. Morgan kaufmann.

[Roberts and Ross, 2010] Roberts, W. R. and Ross, W. (2010). Rhetoric. Cosimo Classics.

[Schmitt, 2002] Schmitt, N. (2002). Applied Linguistics. Hodder & Stoughton. London.

[Searle, 1969] Searle, J. R. (1969). Speech acts: An essay in the philosophy of language.

Cambridge university press.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1{47.

[Sperber, 1987] Sperber, Dan and Deirdre Wilson. (1987) Precis of Relevance: Communication and Cognition. *Behavioral and Brain Sciences.* 10, 697-754.

[Stefanowsky, 2008] Stefanowsky, J. (2008). Multiple classifiers. Catania-Troina.

[Velázquez, 2004] Velázquez, G. (2004). Inprobisazioa in Ahozko inprobisazioa munduan topaketak. Donostia. Euskal Herriko Bertsozale Elkartea. 125-136 or.

[Vidal, 2004] Vidal, M. V. E. (2004). Aportaciones de la pragmatica. Vademecum para la formacion de profesores. Ense~nar espa~nol como segunda lengua (12) 1 lengua extranjera (LE), pages 179{197.

[Verschueren, 1999] Verschueren, J. (1999). *Understanding pragmatics.* London: Arnold.

[Wold et al., 1987] Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1):37{52.

[Zelaia et al., 2005] Zelaia, A., Alegria, I., Arregi, O., and Sierra, B. (2005). Analyzing the efect of dimensionality reduction in document categorization for basque. Archives of Control Sciences, 600:202.

[Zelaia et al., 2011] Zelaia, A., Alegria, I., Arregi, O., and Sierra, B. (2011). A multiclass/multilabel document categorization system: Combining multiple classiers in a reduced dimension. Applied Soft Computing, 11(8):4981{4990.

[Zipitria et al., 2012] Zipitria, I., Sierra, B., Arruarte, A., and Elorriaga, J. A. (2012). Cohesion grading decisions in a summary evaluation environment: A machine learning approach.

7. Laburduren zerrenda

A3	Sujuntiboko adizkiak
ABL	Ablatiboa
ADI	Aditza
ALA	Adlatiboa
BN	Bayes Net (sare bayestarrak)
EDBL	Euskararen datu-base lexikala
GNU	Lizentzia publiko orokorra. Software librea.
HAUL	Hitz anitzeko unitate lexikala
IB1	K-nn algoritmo-sailkatzailea (WEKAn)
INE	Inesiboa
IOR PERARR	Izenordain pertsonal arrunta
IZB	Izen berezia
IZE-ARR	Izen arrunta
K-nn	K nearest neighbor sailkatzailea
LIB	Leku-izen berezia
LOT LOK KONT	Lokailu kontsesiboa
LSA	Laten semantic analisis. Dimentsio murriztailea.
NB	Naive bayes. Algorimo-sailkatzailea
NI_	NORI kasuko aditz morfema
NK_	NORK kasuko aditz morfema
NR_	NOR kasuko aditz morfema
PCA	Principal component analysis. Dimentsio- murriztailea
SMO	Sequential minimal optimization. Algoritmo-sailkatzailea
SVM	Suppor Vector Machine. Algoritmo-sailkatzailea saila.
Tf-Idf	Term frequency–Inverse document frequency.

8. Taulen eta irudien zerrenda

9. Eranskinak

9.1. Programak.

9.1.1. PERL programazio-lengoaia[15]

Larry Wall hizkuntzalariak sortu zuen ikasten zaila ez den sintaxi soil eta malguko lengoaia hau. Testu-fitxategiak izan dira gure azterlaneko euskarria, bada horiek aztertzeko eta manipulatzeko lengoaia gisa sortu zen, berariaz, PERL. Testuok lerroz-lerro irakurtzeko gai da eta karaktereekin zein hitz terminoekin egoki moldatzen da. Erosoa da adierazpen erregularrekin bilaketa eta aldaketa atazetan jarduteko.

Bestalde, software librea izanik eta doan eskuratzeko aukera eskaintzen duenez, erraz lor daiteke zeinahi sistema eragiletan erabiltzeko.

Jarraian, lanean erabili ditugun zenbait programen sarrerako azalpen edo iruzkinak jaso ditugu.

9.1.2. Klase bakoitzarentzat lemma-poltsak sortzekoa

bagWordStop.pl --> zehaztutako klasearentzako bag-of-word sortzen du ARFF formatuan.

```
# Programa: datu-fitxategia, stopwords fitxategia eta klasea emanda, bag-of-word sortzen du klase horrentzako arff formatuan.
#
# Sarrera: 3 argumentu:
#                       - Dat-Fitxategia, bertso-puntuak eta hauei dagozkien klaseekin osatutakoa, formatu honetan:
#                       - Lerro bakoitietan puntuak (bat lerro bakoitzeko)
#                       - Lerro bikoitietan klaseak (zuriunez bananduak)
#                       - StopWords fitxategia
#                       - Zenbaki bat, 1-6 artekoa, klasea adierazten duena

# Irteera: bagLemmaZenb.arff fitxategia, arff formatuan.
#
#
# Erabilera: > perl bagWordStop.pl datBanatuta.txt stopwords.txt zenb
```

15 Astigarraga, A., Gojenola, K., Sarasola, K., Soroa, A. (2009). TAPE testu-analisirako PERL erremintak. UEU.

9.1.3. Lema-poltsa seikotea terminoen Tf-Idf pisaketa aintzat hartuta sortzekoa

bagLemmaStopTfidf --> zehaztutako klasearentzako bag-of-lemma sortzen du (terminoen tf-idf pisaketa) ARFF formatuan.

```
# Programa: datu-fitxategia, stopwords fitxategia eta klasea emanda, bag-of-lemma sortzen du klase horrentzako
#                         tf-idf pisaketarekin arff formatuan.
#
# Sarrera: 3 argumentu:
#                         - Dat-Fitxategia, bertso-puntuak eta hauei dagozkien klaseekin osatutakoa, formatu honetan:
#                                 - Lerro bakoitietan puntuak (bat lerro bakoitzeko)
#                                 - Lerro bikoitietan klaseak (zuriunez bananduak)
#                         - StopWords fitxategia
#                         - Zenbaki bat, 1-6 artekoa, klasea adierazten duena
# Irteera: bagLemmaZenb.arff fitxategia, arff formatuan.
#
#
# Erabilera: > perl bagLemma.pl datBanatuta.txt stopwords.txt zenb
```

9.1.4. Lema-poltsa seikotea erregela morfosintaktikoen arabera

morArffSortu.pl --> Erregeletan oinarritutako ARFF fitxategiak klase bakoitzerako.

```
#       Programa: morfArffSortu.pl
#                               Erregela morfologikoetan oinarritutako arff fitxategiak sortzeko,
#                               1-6, klase bakoitzerako
#
#       Sarrera:
#                       - morfo.txt: atributuen balioak dauzkan fitxategia
#                       - klaseak.txt: puntu bakoitzari dagokion klasea
#                       - klasea: ze klaseri dagokion arff-a, 1-6
#
#       Irteera: klasea-ri dagokion arff fitxategia itzultzen du irteera estandarretik
#
#       Erabilera: perl morfArffSortu.pl morfo.txt klaseak.txt 1
```

9.1.5. Erregelekin lortutako datuak fitxategi erabilgarrian

emaitzakAzt.pl -->Erregela morfologikoekin eraikitako morfo[1-6].txt fitxategiak aztertzen ditu, confusion matrix, performance eta gainerako balioak bueltatuz (WEKAn bezala)

```
#	Programa: emaitzakAzt.pl
#		Erregela morfologikoekin eraikitako morfo[1-6].txt fitxategiak aztertzen ditu,
#		confusion matrix, performance eta gainerako balioak bueltatuz (WEKAn bezala)
#
#	Sarrera:
#
#
#
#	Irteera: azterketaren emaitzak irteera estandarretik
#

#	Erabilera: perl emaitzakAzt.pl
```

9.1.6. Puntuei dagozkien etiketa-fitxategiak sortzekoa (erregela linguistikoen arabera)

morfoAtrib.pl --> Azterketa morfosintaktikoan oinarrituta puntu bakoitzaren atributu zerrenda itzultzen du.

```
# Programa: morfoAtrib.pl
#		Azterketa morfosintaktikoan oinarrituta, puntu bakoitzari dagozkion atributuak (1-6) ezarri
# Sarrera:
#		- puntuen azterketa morfologikoa daukan fitx. (puntuakZatiak.txt)
#		- bertso puntuen fitx. (azken puntua den ala ez zehaztuz) (azkenPunt.txt)
#
# Irteera: morfo.txt. Azterketa morfologikoan oinarrituta, puntu bakoitzari dagozkion atributuak
#		Atributuak: 0 Mezua, 1 Lekua, 2 Publikoa, 3 Saioa, 4 Norbera, 5 Betelana
#
#
# Erabilera: perl morfoAtrib.pl puntuakZatiak.txt azkenPuntu.txt
```

9.1.7. Zeinahi agur-bertso etiketatzeko programa

agurrakEtiketatu.pl --> Agur-bertsoa eta bere azterketa morfosintaktikoa emanda, bertsoa etiketatzen duen programa (erregelak erabiliz).

```
# Programa: agurrakEtiketatu.pl
#                    Agur-bertsoa eta bere azterketa morfosintaktikoa emanda,
#                    bertsoa etiketatzen duen programa
#
# Sarrera:
#                    - Bertsoaren azterketa morfologikoa daukan fitx. (bertsoa.zatiak)
#                    - Bertsoa duen fitx. (bertsoa.txt)
# Irteera: bertsoaEtik.txt. Azterketa morfologikoan oinarrituta, puntu bakoitzari dagozkion atributuak
#                    Atributuak: 0 Mezua, 1 Lekua, 2 Publikoa, 3 Saioa, 4 Norbera, 5 Betelana
#
#
# Erabilera: perl morfoAtrib.pl puntuakZatiak.txt azkenPuntu.txt
```

9.2. Bertso-puntuetan erauzitako edukia (hastapenak)

Lanaren lehen urratsetako berrogei bertsoak eta beraietan antzeman ditugun ezaugarriak (erreferentzia, ideia eta bestelakoak) zerrendatu ditugu ondoko taulan.

Bertsolaria	**Bertsoa**	**Erauzitako edukia**
Millan Telleria 1986-03-09 Bertsolarien txapelketa nagusia. Tolosa Arratsaldeko saioa Hasierako agurra	Talde hontako sei lehen galbaiak hazitako utzi ginun, ta gure zai zeuden biak ere ez datozte oso ilun, ta entzuleak dana gainezka kabitu ezinik inun, eta aurrean gu epaitzeko hona bederatzi lagun; ia guztiak portatzen geran alkar gozatu dezagun. (bis)	Txapelketako ibilbideari erreferentzia. Finaleko kideei. Publikoari. Epaileei. Saio biribila osatzeko itxaropena.
Anjel Mari Peñagarikano 1986-03-09 Bertsolarien txapelketa nagusia. Tolosa Arratsaldeko saioa Hasierako agurra	Estu ta urduri nago ez trankil ta lasai, lanean hasterako noiz amaituko zai, bertsozaleek berriz saio polita nahi, barkatu ni ez naiz gaur horretarako gai, asko ez det egingo bainan al dana bai.	Norbere egoera azaltzen. Estutasuna. Ez dago nahi lukeen bezala, zerbait gertatu zaio eta denek dakite zer. Publikoari erreferentzia. Saio ona osatzeko itxaropena.

Jose Luis Gorrotxategi 1986-03-16 Bertsolarien txapelketa nagusia. Bilbo Goizeko saioa Hasierako agurra	Egun hontan saioa hasi behar lanez, hortarako Bilbora etorri geranez, asmatutzen ditugun haundinak esanez, nahiz eta hortarako jaioak izan ez, agur hau eskeintzen det Bizkaian omenez.	Saioa egiten den lekuari erreferentzia. Apaltasuna. Tokikoei oles. Saio ona osatzeko asmoa.
Millan Telleria 1986-03-16 Bertsolarien txapelketa nagusia. Bilbo Goizeko saioa Hasierako agurra	Lehengoan Beotibar gaurkoan Casilla, Bilbora heldu gera tikearen bila, erteten ez bazaio trenari kurpila, o entzule jatorrak gaur e hainbat mila, bertsoaren sustraia ez daukagu hila.	Txapelketako ibilbideari erreferentzia lekuak aipatuta. Saioa egiten den lekuari erreferentzia. Entzuleei oles. Bertsozaletasunari goratzarre.
Jon Sarasua 1986-03-16 Bertsolarien txapelketa nagusia. Bilbo Goizeko saioa Hasierako agurra	Lehengoan nere marka nuela puskatu, ez pentsa ez nauenik gehiegi juzgatu, bertsolariaz leike horrela prestatu, presio horretatik nahi nuke askatu, eman ezin dugunik guri ez eskatu.	Berak txapelketan egindako ibilbideari erreferentzia. Bertsogintzaz eta txapelketaz iruzkinak. Saioarekiko jarrera erakusten, ez da entzuleen mende jartzen.
Anjel Larrañaga 1986-03-09 Bertsolarien txapelketa nagusia. Tolosa Goizeko saioa Hasierako agurra	Agur ta erdi bertsozaleak lehendabiziko sarreran, behin da berriro jarri gerade kantatutzeko aukeran, ordu ilunak izanagaitik txapelketan gora-beheran, saia nahi degu ta ia zuen gogoko izaten geran.	Entzuleei agurra. Bada gogoan ez dudan zerbait denek ezagun dutena (garai ilunak). Saio ona osatzeko itxaropena.
Luis Otamendi 1986-03-09 Bertsolarien txapelketa nagusia. Tolosa Goizeko saioa Hasierako agurra	Udaberria gainean eta txarrak ez gatoz onduta, eta Tolosa herri honekin zorrik asko badegu ta, Jainkua ere begira degu laino guztiak kenduta, aber zerozer egiten degun berak goitik lagunduta.	Urte sasoia. Saioa egiten den lekuari erreferentzia. Jainkoari erreferentzia (garaian garaiko estandarrak). Saio ona osatzeko itxaropena.
Jose Luis Gorrotxategi 1986-03-09 Tolosa txapelketa. Finalerdia. Arratsaldeko saioa. Hasierako agurra. Gaztalondo handian. Bederatziko handia.	Tolosako frontoian gaur holako gala, zeinek pentsako zuan izango zala, entzuleak txaloak lasai jo ditzala, sinistu gu horrekin poztutzen gerala, eguna etortzean txoriak bezala. (bis)	Tokiari erreferentzia. Entzuleei diosala eta babes eskaera.

Mikel Mendizabal 1986-03-09 Tolosa txapelketa. Goizeko saioa. Hasierako agurra. Loreak udan ihintza bezala. Zortziko handia.	Utzi ditzagun kontuak eta hartu ditzagun arretak, bero faltarik ez daukate gaur Beotibarko paretak, oztopo asko baditu ere bertsoaren neurketak, giro aldetik mesede asko badakarzki txapelketak.	Lekuari erreferentzia. Entzuleek eskainitako berotasunari loreak. Txapelketari erreferentzia.
Imanol Lazkano 1989-12-17 Donostia txapelketa. Hasierako agurra. Haize hegoak hurbiltzen ditu. Hamarreko handia.	Anai-arreba bihotzekoak egunon bana lehenbizi, zein polita dan bertsoa maite eta bertsoetan bizi, gaur sufritzera gatozte baina dena ez al du merezi? San Antonioi piztu dizkiot sei kandela ta bi zuzi, halare hemen Jesukristonak beharko dira ikusi. (bis)	Entzuleei agurra. Bertsolaritzari goratzarrea. Txapelketak eragiten dituen estutasunak eta pozak. Saio onaren bila babes eske; santuak, jainkoa.
Nikolas Zendoia 1989-12-03 Gernika Lumo. Txapelketa 11:30 Hasierako agurra Uso xuria errazu	Bihotza dut zabaltzeko egunonak eskeintzeko, (bis) prestu nago ni bertsotarako ez Donostira sartzeko, nere azkenengo itxaperoak Gernika zaharre(a)n lehertzeko. (bis)	Egun on, arratsalde on... adeitasuneko agur formak. Apaltasun adierazpenak. Saio ona eskaintzeko itxaropena. Saioa egiten den lekuarekiko erreferentzia.
Bittor Elizagoien 1990-02-24 Nafarroako bertsolari txapelketa (Ituren) Hasierako agurra Hamarreko handia	Usteko zue lasai nagola baina nago apuruan nere lagunek esan didate zerbait ein zak aurtenguan ta andregaiak ez nauela admetituko onduan ta amak berriz joko nauela isetsarekin buruan baldin azkena gelditzen banaiz etxera ezin naiz juan.	Norbere estutasuna agertu eta babesa eskatu. Lagunei eta jarraitzaileei erreferentzia. Presioaren berri ematea (andregaia, ama...).
Manolo Arozena 1990-04-01 Nafarroako bertsolari txapelketa (Lekunberri) Hasierako agurra Hamarreko handia	Arratsaldeon lehendabiziko nere arreba ta anai bi talde gatoz bertso kantari Taberna harturik artzai norbait alperrik egon liteke bertso sakonen batzun zai Nafarroako bertsolaritza dugu helburu eta gai bizirik dela erakustea besterik ez genduke nahi.	Adeitasuneko agur formulak: arratsalde on... Gai jartzaileari erreferentzia. Lanari buruzko azalpena (zer egin behar duten). Bertsolaritzaren alde, (Nafarroako txapelketan, bertakoari bultzaka).
Ireneo Ajuria 1990-12-23 Araba eta Bizkaiko bertsolari txapelketa (Bilbo) Bukaerako agurra Arratsalde on lehendabizikotik Zortziko handia	Beti penatan ez da izaten euskaldunaren bizitza toki honetan txaloak eta odolak goitik dabiltza baina azkenean emango dute gaurko txapelan emaitza mila esker ta sendatu daigun Bizkaiko bertsolaritza (bir)	Zorionaren aldarria (zein ederra den elkarrekin biltzea, anaitasuna...) Azkenean egitekoaren larria (txapela, lan egokia...) Bertsolaritzaren alde (Bizkaiko txapelketan, bertakoari bultzaka).

Manolo Arozena 1991-03-07 Nafarroako bertsolari txapelketa (Doneztebe) Hasierako agurra. Zortziko handia	Bost kilometro oinez egina beraz, ni ez nator hotzez nerbio eta txorradak utzi ta kanta nahi det bihotzez kategoririk ez detela ta zertako isildu lotsez? hau ikusita urte bete bat behintzat pasako det pozez.	Norbere egoera azaldu eta lanerako prestutasuna agertu. Lotsak baztertuta bere neurria ematera. Jendeari eskerrona saiora bertaratzeagatik.
Jexux Arzallus 1993-12-05 Euskal Herriko txapelketa nagusia (Eibar) Hasierako agurra Santa Barbara zure bizitza Hamarreko handia	Hendaian sartu kotxean eta bi arnasakin Eibarra. Ezin antzeman nun hasi eta nun bukatzen zen dardarra. Pauso hau ere egun batean zegoen eman beharra; hauxen da orain buru barruan dabilen kezka bakarra: orain hegoak baditut baina ez dut galdu nahi iparra.(bis)	Norbere jatorria. Saioa egiten den tokiari buruzko erreferentzia. Norbere egoeraren adierazpena (estutasuna, larritasuna...) Bertsolariak aurrena duen erronkaren neurria eta bere helburua.
Unai Iturriaga 1993-12-05 Euskal Herriko txapelketa nagusia (Eibar) Bukaerako agurra Amodioa gauza tristea Hamarreko handia	Bukatu nahi dut nere jertseak esaten duen bezela, gu hemen pozik bagaude ere hortxe baitago kartzela, eta badakit hamabost urtez jentea hor dagoela. Nere partetik beraientzako omen ta txalo itzela, telebistatik beraien berri ez baita izango bestela. (bis)	Unean uneko elementuei egindako erreferentziak (garaiko albiste eta berriak, honakoan Iturriagak zeraman jertseko aldarriari buruzkoak). Kritika soziala. Kartzela. Saioan izan ez daitezkeenentzako omenaldia.
Aritz Lopetegi 1993-12-19 Euskal Herriko Txapelketa nagusia (Donostia) Arratsaldeko saioa (17:00) Hasierako agurra Mutil koxkor bat itsuaurreko Hamarreko handia	Agur honekin Leioa aldean jarri nahi nuke arreta; kontratu bako irakasleak dira, hain zuzen, nere meta. Egunik egun atean daude hotzetik ezin gordeta; besarkada bat bialtzen diet elkartasunez beteta, beste guztiek batera baino gehio balio dute ta. (bis)	Unean uneko elementuei egindako erreferentziak (garaiko albiste eta berriak, honakoan EHUko irakasle kontratu gabeak). Kritika soziala. Laneko egonkortasunik eza, baldintza duinak. Haientzako omenaldia.
Jokin Sorozabal 1997-11-02 Euskal Herriko Txapelketa nagusia (Zarautz) Hasierako agurra Ai gure antzinako Hamarreko txikia	Kantuz hasi baino lehen erantzuna txaloz; (bis) frontoia bete zaigu hintxadaz ta morboz. Zazpirehun puntuetan listoia apropos, gure intentzioa hobea da askoz: iristera ez baina egitera gatoz. (bis)	Bertsozaleentzako eskerrona eskainitako babesagatik. Testuinguruaren deskribapena. Norbere asmoak, apaltasunetik gogor saiatzea.
Maialen Lujanbio 1997-11-16 Euskal Herriko Txapelketa nagusia (Bera) Hasierako agurra Anderea gorarik Zortziko txikia	Bertsolai eta entzule gaurkoz bat dia; hainbeste txalo nola liteke ordia? Hola egin leike bertso saioen bidia, zuek erdia eta guk beste erdia.	Tranpaldoaren bi aldeetakoak bat egin nahi. Batasunaren irudia. Eskerrona txalo eta babesagatik. Entzuleen garrantzia goraipatu, erdibanako lana.

Jexux Mari Irazu 1997-11-02 Euskal Herriko Txapelketa nagusia (Zarautz) Hasierako agurra Saltarina da txepetxa eta Zortziko handia	Bertsota ezin etorri leike urduritasunak jua; patxara berriz izan liteke komeni ez dan lujua. Beti bezela gaur'e tenplian egon liteke gakua, bertsolaritzan bi aurpegiren islada baita gaurkua: jolasa eta jokua.	Norbere egoeraren berri ematea. Lanerako behar dena: Erabakimena, patxada, tenplea... Bertsolaritzaren definizio bitarra: jolasa eta jokoa.
Jon Maia 1997-11-16 Euskal Herriko Txapelketa nagusia (Bera) Hasierako agurra Iparragirre abila dela Neurri berezia	Udazkeneko tonu goxoak duen kolore gordina, nola liteke infernuraino horren bide atsegina? Ohorezko borroka, hau da gure erronka, ia nor den izokina, Bidasoatik Urumeara salto egiten dakina. (bis)	Urte sasoiari buruzko erreferentzia. Saioak eragiten duen larritasun goxoaren inguruko irudia. Lekuari eta egoerari erreferentzia (ibaian gora doan izokinarena)
Jon Martin 2001-03-23 Euskal Herriko Txapelketa nagusia. Gipuzkoako sailkapena (Pasaia). Hasierako agurra Baserrian jaio nintzan	Bianditzen sortu eta, beste errekekin elkartu, ondoren Oiartzuneko auzo askotan pasatu, Arraguan jira egin ta Orereta zeharkatu; Oiartzun ibaia ohi da Pasaiara ailegatu. Ur ttantta batek hasten du ta itsasoak bukatu, zentzu bakarrean doa nahiz ta bost zentzu goxatu. Ta nik ere ibai horren parte nahi dut bilakatu, eta zuen belarritan nahi nuke itsasoratu	Saioa egiten den lekuari erreferentzia (eskualdeko herriak zeharkatzen dituen ibaiaren ibilgua baliatuta) Nondik datorren bera, nora etorri den. Batasunaren goratzarrea, entzuleekiko elkartasuna (elkarlana, bultzada, denok batera).
Amets Arzallus 2001-04-01 Nafarroako bertsolari txapelketa (Bera) **Txapeldunaren agurra** Baserrian jaio nintzan	Txapel hau zuzendu nahi dut hainbat pertsonarengana: lehendabizi Joxe Mari, bide erakusle dana. Bigarrenik futboleko talde ta lagunengana. Nola ez bertso eskola, bi arrebak eta ama! Eta beste eskeintza bat maitasun osoz doana: gehien lagundu didaten Elorri eta Joana. Eta azkenikan aita, dena doa zuregana, zuk eman didazulako naizena ta daukadana.	Txapelaren eskaintza. Lagun eta euskarri izan dituenak gogoan.
Fredi Paia 2001-12-02 Euskal Herriko bertsolari txapelketa nagusia (Tolosa) Hasierako agurra Zibilak esan naute Zortziko txikia	Lagunek esan naute "kantatu egoki", Tolosan behar dala gauza erabaki. Sailkatzeko behar nituzke beatzirehun ta zazpi, barre egingo luke nere amak baleki.	Saioa egiten den lekuari erreferentzia (hitzetan eta baliatutako doinuan). Txapelketaren nondik norakoen azalpena (behar diren puntuak, sailkapena...). Etxekoak eta lagunak gogoan.

Aitor Sarriegi 2003-11-23 Gipuzkoako bertsolari txapelketa (Zarautz) Hasierako agurra Hamarreko txikia	Orain seiren bat urte hementxe juntatu ta final-laurden baten genuen kantatu. Urteak pasa dira ezin da ukatu baina finalerarte hau ez da bukatu urteak pasa dira ezin da ukatu ea danak ez diren alperrik pasatu.	Aurrekariak gogoratzen ditu. Txapelketaren nondik norakoei erreferentzia (zer egin behar den, nola osatu ibilbidea). Norbere asmoak eta itxaropena.
Aitor Mendiluze 2005-12-18 Euskal Herriko txapelketa nagusia (Barakaldo) Goizeko saioa 11:00 Bukaerako agurra Txikitatikan edukia dut Neurri berezia	Mila zorion Andoni zuri lau badituzu jasoak besteengatik ere astindu nahi ditut nere besoak. Bueno nik zerbait egin det ondo izan arren arazoak igual ez ziren goxoak ta entzuteko erosoak baina barruak eskatzen zidan ta bota ditut osoak orain bi urte kantatu gabe utzi nituen bertsoak. (bis)	Txapeldunari aitorpena, baina baita gainerako bertsolariei ere. Egindako lanaren laburpena. Antzeko beste egoeratan bizitakoari erreferentzia (bi urte lehenago egin ezaren damua, oraingoarekin estalia).
Ainhoa Agirreazaldegi 2009-11-14 Euskal herriko txapelketa nagusia (Tolosa) Bukaerako agurra Larogeita hamar urte Hamaseiko berdina, 8 puntuz eta 8 silabaz	Elkarteko lagun denei nere esker zintzoenak orain bi urte honekin lanean hasi zirenak aspaldi erloju gabe bizitzen ohitu direnak jeikitzen lehenak eta erretiratzen azkenak (bis) a ze entrenamentuak epaile taldearenak! bilera amaitezinak dira gai-jartzaileenak... Nola ahaztu herriz herri laguntzen gaituzten denak izarrek argi egiteko zerua sortzen dutenak. (bis)	Eskaintza lagun eta babes izan dituenei (txapelketa antolatzen aritutakoei, epaileei, gai-jartzaileei...) **Izarrek argi egiteko zerua sortzen dutenei**
Igor Elortza 2009-11-21 Euskal herriko txapelketa nagusia (Gernika-Lumo) Bukaerako agurra Nere gorputza dardarka daukat Hamaseikoa, handiaren moldekoa, 9 puntuz	Nola ibilbide bat egiten den mantso pausotik pausora Santi ta Josun herriak ere behar du bere denbora. Taldekeriak utzi ta bildu batzuk besten abarora boto guztiak udaletara preso guztiak kanpora estatu bila biolentzia bako konfrontaziora. Ta bertsoak berriz nora? Oraingoz Barakaldora. Dudarik barik pozik igoko nintzatekeen arren gora zuotako bat izatea ere ohore bat izango da.	Saioa egiten den lekuari erreferentzia. Gizarte gaiak (Elkartasuna, batasuna): Hauteskundeak, presoak, herria, borroka zibila. Txapelketaren ibilbidea eta bere egoera.

Uxue Alberdi 2009-11-29 Euskal herriko txapelketa nagusia (Donostia) Hasierako agurra Gure herri ttiki hau da Hamaseiko berdina, 8 puntuz eta 8 silabaz	Izate batek berezko izaten du kontrastea gaur arte erabaki dut alde bat lehenestea nire aurpegi sentibera goxoa aberastea eta plazerra izan da xuxurlaka abestea baina badaukat letxea izatearen ospea badauzkat errebeldia eta amorru askea. Gaur erakutsi nahi ditut alde bat eta bestea nahi nuke biak sentitu eta sentiaraztea. (bis)	Bertsolariaren izaeraren berri eman eta saiorako duen asmoaren azaltzen da. Ordura artekoak gogoan izanik itxura berria ere erakutsi nahia.
Maialen Lujanbio 2009-12-13 Txapelketa nagusia (Barakaldo) Arratsaldeko saioa 16:30 Bukaerako agurra Eguzkiak urtzen du han goian A Hamaseiko handia	Gogoratzen naiz lehengo amonen zapi gaineko gobaraz gogoratzen naiz lehengo amonaz gaurko amaz ta alabaz. Joxei ta zuei mila zorion miresmenaren zirraraz ta amaituko dut txapel zati bat zuek guztiontzat lagaz. Gure bidea ez da errexa bete legez, juizioz, trabaz... Euskal Herriko lau ertzetara itzuliko gara gabaz eta hemen bildu dan indarraz grinaz eta poz taupadaz herri hau sortzen segi dezagun euskaratik ta euskaraz.	Txapela eskaintzea: emakumeei (zahar, heldu eta gazte), lagunei eta ikusleei. Gizarte gaiak: Euskal Herria eta euskara. Elkartasuna, batasuna: euskaratik euskaraz.
Sustrai Colina 2009-12-13 Txapelketa nagusia (Barakaldo) Goizeko saioa 11:00 Hasierako agurra Nere gorputza dardarka daukat Hamaseikoa, handiaren moldekoa, 9 puntuz	Nahiz eta kirol nazionala izan pronostikogintza Patxi Lopezek ustekabean lortu zun Lehendakaritza ta Gonzalezek lau ta erdiko txapel handiaren ditxa nork iragarri zuen ETBn Erregearen espitxa? Nork Barakaldo izango zela bertsoaren bihotz mintza? Bertsozale, herrigintza kaxo zer moduz gabiltza? Sorpresa txikiz ezuste handiz beteta dago bizitza ta ez nintzateke lasai egongo faboritoa banintza.	Unean uneko elementuei egindako erreferentziak (garaiko albiste eta berriak, Patxi Lopezen lehendakaritza, Gonzalezen 4 1/2ko txapela, errege ETBn...). Txapelketari erreferentzia (ustekabekoak ugari izanik, faboritoak ez beude lasai).
Jon Lopategi 1982-11-28 Txapelketa nagusia (Idiazabal) Bukaerako agurra Zortziko handia	Hor joan dira gure kantuak deiturak eta izenak berso guziak ezin atera dezio bezain zuzenak baina gehiago nola ez duen gaurkoz gure almazenak suerte on bat izan zatela arratsaldez datozenak.	Egindako lanaren errepasoa. Okerren damua eta ahaleginaren aitorpena. Hurrengo saiokoei onena opa die.

Txomin Garmendia 1982-12-05 Txapelketa nagusia (Durango) Hasierako agurra Zortziko handia	Gure hizkuntzik ederrenean kantuan bersolariak zuen gogoko izango al dira hemengo berso berriak, irripar bat ta alaitasun bat premizko dauka erriak hasierako besarkada bat entzule maitagarriak.	Bertsolarien asmoen berri. Ahaleginik beteena. Apaltasuna. Gizarte gaiak. Herria triste, alaitasunaren aldeko jarrera. Entzuleak beti gogoan.
Joxe Mari Altuna 1982-12-05 Txapelketa nagusia (Durango) Arratsaldeko saioa 17:00 Hasierako agurra Zortziko handia	Motibo planko bada etortzeko orain Durango aldera lehendabiziko nijoakizue atsaldeon ematera nere burutik gauza haundirik ez dakit leiken atera al dedan dana egitera nator lagun hauekin batera.	Saiora biltzearen inguruko arrazoiak. Adeitasuneko formulak: arratsalde on. Apaltasuna. Ahalegina eskaini talde-lanean aritzeko.
Juan Mari Narbaiza 1986-03-09 Txapelketa nagusia (Tolosa) Arratsaldeko saioa 17:00 Bukaerako agurra Praixku Galtzarreta A Zortziko txikia	Tolosa portatu da eskertzen det hori, naiz ta nerbioekin ia ia erori, hurrengoz kantatu nahi diogu Bilbori, kandela bana piztu San Antoniori.	Saioa egiten den lekuari buruzko erreferentzia eta harekiko eskerrona. Norbere egoeraren berri ematea: estu, larri... Txapelketan aurrera egiteko asmoa (Bilbon kantatu) Babes eskaera hango santuei (San Antonio).
Iñaki Murua 1986-03-09 Txapelketa nagusia (Tolosa) Hasierako agurra Mutil koxkor bat itsuaurreko Hamarreko handia	Zuentzat nere aitormen fina bihotzaren barreneti, eta Aitzoli omen eder bat egin zuenarengati, itzultzaileen eskolakoei morala eman galanki, zuek eta gu ta hoiek ere gabiltza bide bereti, gure hizkuntza indartu nahirik gauza guzien gaineti.	Entzuleak gogoan, agurra eta eskertza. Omenaldia bidegileei eta oinarri jartzaileei (Aitzol). Euskararen aldeko aldarria (denok batera bultza dezagun).
Anjel Larrañaga 1986-03-16 Txapelketa nagusia (Bilbo) Bukaerako agurra Zortziko handia	Ez dakit nola aterako dan herriaren kiniela, baino badakit batzuak behintzat burruka latza dutela, datorren jaian ez dakit zeinek jantziko duen txapela, nahi deguna da saio jator bat eskeini dezaigutela.	Txapelketaren nondik norakoez hausnarketa. Borroka estua txapela janzteko lehian. Bertsolariaren asmoa, batik bat saio ederra ateratzea.
Jon Sarasua 1986-03-23 Txapelketa nagusia (Donostia) Bukaerako agurra Hamarreko txikia	Danetan gazteena nerau naizelako, denen partez agurra asmatu beharko, hornitu gabe dena egoten da plako, goizean bota degu kaloria franko, berriz bildu ditzagun arratsalderako.	Norbere egoeraren berri eman. Jendeari agur abegikorra, arina, dibertigarria, atsegina eta bazkaltzerako deia. Hurrengo saiorako asmoak.
Mattin 1962-12-30 Txapelketa nagusia (Donostia) Bukaerako agurra (Zazpigarren postuan geratu zen Mattinen agurra)	Azkena kantatzeko orai erran gaituzte, partitzian nik guziak eskertu nai nituzke, zar eta gazte, gizon ta emazte, bertze aldi bat arte, deneri goraintzi eta ondo bizi zaitezte.	Agur formala. Ikusle denei eskerrak eta hurrengo baterako gonbita. Eskerrak. Zahar, gazte, gizon, emazte. Hurrengora arte. Goraintziak. Izan ongi.

Towards Basque Oral Poetry Analysis: A Machine Learning Approach

Mikel Osinalde, Aitzol Astigarraga, Igor Rodriguez and **Manex Agirrezabal**
Computer Science and Artificial Intelligence Department,
University of the Basque Country (UPV/EHU), 20018 Donostia
teagenes@hotmail.com
aitzol.astigarraga@ehu.es
igor.rodriguez@ehu.es
manex.agirrezabal@ehu.es

Abstract

This work aims to study the narrative structure of Basque greeting verses from a text classification approach. We propose a set of thematic categories for the correct classification of verses, and then, use those categories to analyse the verses based on Machine Learning techniques. Classification methods such as Naive Bayes, k-NN, Support Vector Machines and Decision Tree Learner have been selected. Dimensionality reduction techniques have been applied in order to reduce the term space. The results shown by the experiments give an indication of the suitability of the proposed approach for the task at hands.

1 Introduction

Automated text categorization, the assignment of text documents to one or more predefined categories according to their content, is an important application and research topic due to the amount of text documents that we have to deal with every day. The predominant approach to this problem is based on Machine Learning (ML) methods, where classifiers learn automatically the characteristics of the categories from a set of previously classified texts (Sebastiani, 2002).

The task of constructing a document classifier does not differ so much from other ML tasks, and a number of approaches have been proposed in the literature. According to Cardoso-Cachopo and Oliveira (2003) , they mainly differ on how documents are represented and how each document is assigned to the correct categories. Thus, both steps, document representation and selection of the classification method are crucial for the overall success. A particular approach can be more suitable for a particular task, with a specific data, while another one can be better in a different scenario (Zelaia et al., 2005; Kim et al., 2002; Joachims, 1998).

In this paper we analyse the categorization of traditional Basque impromptu greeting verses. The goal of our research is twofold: on the one hand, we want to extract the narrative structure of an improvised Basque verse; and, on the other hand, we want to study to what extent such an analysis can be addressed through learning algorithms.

The work presented in this article is organized as follows: first we introduce Basque language and *Bertsolaritza*, Basque improvised context poetry, for a better insight of the task at hand. Next, we give a general review of computational pragmatics and text classification domains, examining discourse pattern, document representation, feature reduction and classification algorithms. Afterwards, the experimental set-up is introduced in detail; and, in the next section, experimental results are shown and discussed. Finally, we present some conclusions and guidelines for future work.

2 Some Words about Basque Language and *Bertsolaritza*

Basque, *euskara*, is the language of the inhabitants of the Basque Country. It has a speech community of about 700,000 people, around 25% of the total population. Seven provinces compose the territory, four of them inside the Spanish state and three inside the French state.

Bertsolaritza, Basque improvised contest poetry, is one of the manifestations of traditional Basque culture that is still very much alive. Events and competitions in which improvised verses, *bertso*-s, are composed are very common. In such performances, one or more verse-makers, named *bertsolaris*, produce impromptu compositions about topics or prompts which are given to them by a theme-prompter. Then, the verse-

maker takes a few seconds, usually less than a minute, to compose a poem along the pattern of a prescribed verse-form that also involves a rhyme scheme. Melodies are chosen from among hundreds of tunes.

Figure 1: *Bertsolari Txapelketa Nagusia*, the national championship of the Basque improvised contest poetry, held in 2009

When constructing an improvised verse strict constraints of meter and rhyme must be followed. For example, in the case of a metric structure of verses known as *Zortziko Txikia* (small of eight), the poem must have eight lines. The union of each odd line with the next even line, form a strophe. And each strophe, in turn, must rhyme with the others. But the true quality of the *bertso* does not only depend on those demanding technical requirements. The real value of the *bertso* resides on its dialectical, rhetorical and poetical value. Thus, a *bertsolari* must be able to express a variety of ideas and thoughts in an original way while dealing with the mentioned technical constraints.

The most demanding performance of Basque oral poetry, is the *Bertsolari Txapelketa*, the national championship of *bertsolaritza*, celebrated every four years (see Fig.1). The championship is composed by several tasks or contests of different nature that need to be fulfilled by the participants. It always begins with extemporaneous improvisations of greetings, a first verse called *Agurra*. This verse is the only one in which the poet can express directly what she/he wants. For the rest of the contest, the theme-prompter will prescribe a topic which serves as a prompt for the *bertso*, and also the verse metric and the number of iterations. For that reason, we thought the *Agurra* was of particular interest to analyse ways verse-makers use to structure their narration.

3 Related Work

3.1 Computational Pragmatics

As stated in the introduction, the aim of this paper is to notice if there is any discourse pattern in greeting verses. In other words, we are searching certain defined ways verse-improvisers in general use to structure their discourse.

If the study of the meaning is made taking into account the context, we will have more options for getting information of the factors surrounding improvisation (references, inferences, what improvisers are saying, thinking, self-state, context). The field that studies the ways in which context contributes to meaning is called pragmatics. From a general perspective, Pragmatics refers to the speaker and the environment (Searle, 1969; Austin, 1975; Vidal, 2004).

The study of extra-linguistic information searched by pragmatics is essential for a complete understanding of an improvised verse. In fact, the understanding of the text of each paragraph does not give us the key for the overall meaning of the verse. There is also a particular world's vision and a frame of reference shared with the public; and, indeed, we have been looking for those keys. We believe that the verse texts are not linear sequences of sentences, they are placed regarding a criterion and the research presented here aims to detect this intent.

Therefore, searching for the discourse facts in greeting verses led us to study their references.

3.2 Text Categorization

The goal of text categorization methods is to associate one or more of a predefined set of categories to a given document. An excellent review of text classification domain can be found in (Sebastiani, 2002).

It is widely accepted that how documents are represented influences the overall quality of the classification results (Leopold and Kindermann, 2002). Usually, each document is represented by an array of words. The set of all words of the training documents is called vocabulary, or dictionary. Thus, each document can be represented as a vector with one component corresponding to each term in the vocabulary, along with the number that represents how many times the word appears in the document (zero value if the term does not occur). This document representation is called the bag-of-words model. The major drawback of this

text representation model is that the number of features in the corpus can be considerable, and thus, intractable for some learning algorithms.

Therefore, methods for dimension reduction are required. There exists two different ways to carry out this reduction: data can be pre-processed, i.e., some filters can be applied to control the size of the system's vocabulary. And, on the other hand, dimensionality reduction techniques can be applied.

3.2.1 Pre-processing the Data

We represented the documents based on the aforementioned bag-of-word model. But not all the words that appear in a document are significant for text classification task. Normally, a pre-processing step is required to reduce the dimensionality of the corpus and, also, to unify the data in a way it improves performance.

In this work, we applied the following pre-processing filters:

- **Stemming**: remove words with the same stem, keeping the most common among them. Due to its inflectional morphology, in Basque language a given word lemma makes many different word forms. A brief morphological description of Basque can be found in (Alegria et al., 1996). For example, the lemma *etxe* (house) forms the inflections *etxea* (the house), *etxeak* (houses or the houses), *etxeari* (to the house), etc. This means that if we use the exact given word to calculate term weighting, we will loose the similarities between all the inflections of that word. Therefore, we use a stemmer, which is based on the morphological description of Basque to find and use the lemmas of the given words in the term dictionary (Ezeiza et al., 1998).

- **Stopwords**: eliminate non-relevant words, such as articles, conjunctions and auxiliary verbs. A list containing the most frecuent words used in Basque poetry has been used to create the stopword list.

3.2.2 Dimensionality Reduction

Dimensionality reduction is a usual step in many text classification problems, that involves transforming the actual set of attributes into a shorter, and hopefully, more predictive one. There exists two ways to reduce dimensionality:

- **Feature selection** is used to reduce the dimensionality of the corpus removing features that are considered non-relevant for the classification task (Forman, 2003). The most well-known methods include: Information Gain, Chi-square and Gain Ratio (Zipitria et al., 2012).

- **Feature transformation** maps the original list of attributes onto a new, more compact one. Two well-known methods for feature transformation are: Principal Component Analysis (PCA) (Wold et al., 1987) and Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Hofmann, 2001).

The major difference between both approaches is that feature selection selects a subset from the original set of attributes, and feature transformation transforms them into new ones. The latter can affect our ability to understand the results, as transformed attributes can show good performance but little meaningful information.

3.2.3 Learning Algorithms

Once the text is properly represented, ML algorithms can be applied. Many text classifiers have been proposed and tested in literature using ML techniques (Sebastiani, 2002), but text categorization is still an active area of research, mainly because there is not a general faultless approach.

For the work presented here, we used the following algorithms: Nearest Neighbour Classifier (IBk) (Dasarathy, 1991), Nave Bayes Classifier (NB) (Minsky, 1961), J48 Decision Tree Learner (Hall et al., 2009) and SMO Support Vector Machine (Joachims, 1998).

All the experiments were performed using the Weka open-source implementation (Hall et al., 2009). Weka is written in Java and is freely available from its website [1].

In Fig.2, the graphical representation of the overall Text Classification process is shown.

4 Experimental Setup

The aim of this section is to describe the document collection used in our experiments and to give an account of the stemming, stopword deletion and dimensionality reduction techniques we have applied.

[1] http://www.cs.waikato.ac.nz/ml/weka/

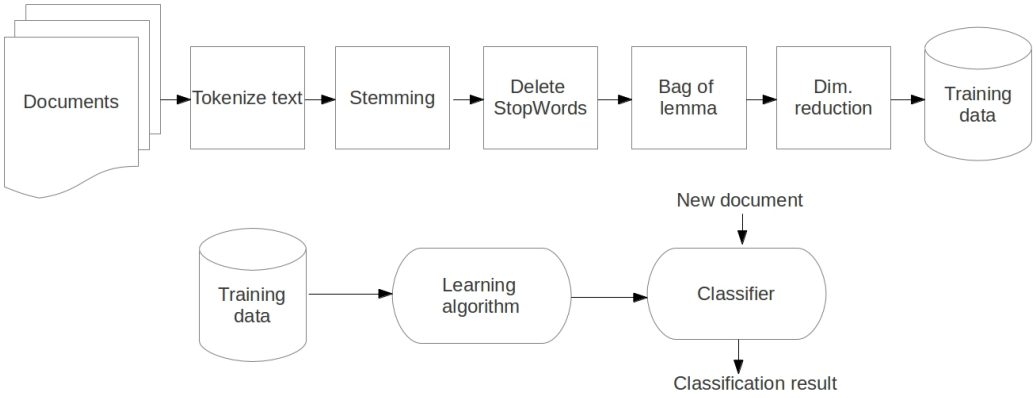


Figure 2: The overall process of text categorization

4.1 Categorization

To make a correct categorization of the verses, before anything else the unit to be studied needs to be decided. We could take as a unit of study the word, the strophes or the entire verses. Considering that we want to extract the structure that would provide information about the decisions made by the improviser and the discourse organization, we decided that the strophe[2] was the most appropriate unit to observe those ideas. Therefore, the first job was to divide the verses in strophes. After that, we began to identify the contents and features in them. The goal was to make the widest possible characterization and, at the same time, select the most accurate list of attributes that would make the strophes as much distinguishable as possible.

We sampled some strophes from the verse corpus described in section 4.2 and analysed them one by one. We had two options when categorizing the strophes: first, analyse and group all the perceived topics, allowing us to propose a realistic classification of the strophes from any verse. And second, make a hypothesis and adjust the obtained data to the hypothesis. We decided to take both paths.

After analysing each of the strophes and extracting their topics, we made the final list, sorted by the relevance of the categories. We obtained a very large list of contents and we arranged it by the importance and by the number of appearance. But that thick list did not help us in our mission as we wanted. So we agreed to try to define and limit the collection of attributes. And we decided to use the second option. Therefore, we studied the foundations of discourse analysis (Roberts and Ross, 2010; Gumperz, 1982), and the classifications proposed by critics of the improvisation field (Egaña et al., 2004; Diaz Pimienta, 2001); and then, we compared them with our predicted one. Merging both approaches we tried to build a strong set of categories.

Combining inductive and deductive paths we formed a list of six categories. So the initial big list that we gathered was filtered to a more selective classification. Therewith, we found possible to label the majority of the strophes in the analysed verses, and also get a significant level of accuracy.

Thus, these are the categories to be considered in the verse classification step:

1. Message: the main idea
2. Location: references to the event site
3. Public: messages and references relating to the audience
4. Event: messages and references relating to the performance itself
5. Oneself aim or Oneself state
6. Miscellaneous: padding, junk. Sentences with no specific meaning or intend.

As well as the five categories closely linked to the communication situation, there is another that we called Miscellaneous (padding, filling). Due to

[2] a pair of stanzas of alternating form on which the structure of a given poem is based

the demanding nature of the improvisation performances, they usually are sentences not very full of content and intent.

We have decided to consider each one of them as a separate goal, and hence six classifiers were to be obtained, one for each category. Thus, each categorization task was addressed as a binary classification problem, in which each document must be classified as being part of $category_i$ or not (for example, Location vs. no Location).

4.2 Document Collection

For the task in hands, we decided to limit our essay to greeting verses from tournaments. We selected 40 verses of a corpus of 2002 verses and divided them into strophes (212 in total). But when we began assigning categories (1-6) to each strophe, we realized we were in blurred fields. It was pretty difficult to perform that task accurately and we thought it was necessary to ask some expert for help. Mikel Aizpurua[3] and Karlos Aizpurua[4] (a well-known judge the former and verse improviser and Basque poetry researcher the latter) agreed to participate in our research, and they manually labelled one by one the 212 strophes.

In that study, we considered each binary class decision as a distinct classification task, where each document was tested as belonging or not to each category. Thus, the same sentence could effectively belong to more than one categories (1 to 6 category labels could be assigned to the same sentence).

As an example, let us have a look to an initial greeting verse composed by Anjel Larrañaga, a famous verse-maker (see Fig.3).

There we can see that each strophe (composed of two lines), was labelled in one, two or even tree different categories.

- (1) (3): Message, Public
- (5): Oneself aim
- (4) (5): Event, Oneself state
- (1) (5) (3): Message, Oneself aim, Public

The document categorization process was accomplished in two steps: during the training step, a general inductive process automatically built a classifier by learning from a set of labelled documents. And during the test step, the performance of the classifier was measured. Due to the small size of our manually categorized corpus, we used the k-fold cross-validation method, with a fold value of k=10.

Agur ta erdi bertsozaleak
*lehendabiziko **sarreran**,*
behin da berriro jarri gerade
*kantatutzeko **aukeran**,*
ordu ilunak izanagaitik
*txapelketan gora-**beheran**,*
saia nahi degu ta ia zuen
*gogoko izaten **geran**.*

As a first introduction,
greetings to all improvisation fans. (1) (3)
Many times we were ready
to sing like now! (5)
Even though there are hard times
in our championship contest, (4) (5)
We will try to make our best
and we hope you find it to your liking! (1) (5)
(3)

Figure 3: A welcome verse composed by Anjel Larrañaga

4.3 Pre-processing the Data

In order to reduce the dimensionality of the corpus, two pre-processing filters were applied. On the one hand, a stopword list was used to eliminate non-relevant words. On the other hand, a stemmer was used to reduce the number of attributes.

The number of different features in the unprocessed set of documents was 851, from which were extracted 614 different stems and 582 terms after eliminating the stopwords. So finally, we obtained a bag-of-lemmas with 582 different terms.

5 Experimental Results

In this section we show the results obtained in the experiments. There are various methods to determine algorithms' effectiveness, but precision and recall are the most frequently used ones.

It must be said that a number of studies on feature selection focused on performance. But in many cases, as happened to us, the are few in-

[3] http://bdb.bertsozale.com/en/web/haitzondo/view/-Mikel-Aizpurua

[4] http://bdb.bertsozale.com/en/web/haitzondo/view/-Karlos-Aizpurua

Category	ML method	Attribute selection	Performance	F-measure
Message	1-nn	None	64.62%	0.62
Location	SMO	InfoGain	89.62%	0.86
Public	SMO	ChiSquare	83.01%	0.81
Event	5-nn	None	78.30%	0.76
Oneself	SMO	InfoGain	62.26%	0.60
Miscellaneous	1-nn	GainRatio	87.74%	0.83

Table 1: Best results for each category

stances of positive classes in the testing database. This can mask the classifiers performance evaluation. For instance, in our testing database only 22 out of 212 instances correspond to class 2 ("Location"), giving an performance of 90.045 % to the algorithm that always classifies instances as 0, and thereby compressing the range of interesting values to the remaining 9.954 %. Therefore, in text categorization tasks is preferred the F-measure, the harmonic average between precision and recall.

Table1 shows the configurations that have achieved the best results for each category.

Based on the results of the table, we can state that they were good in three out of six categories (Location, Public and Miscellaneous); quite acceptable in one of them (Event); and finally, in the remaining two categories (Message and Oneself) the results were not very satisfactory.

Regarding to the learning algorithms, it should be pointed out that SMO and k-nn have shown the best results. We can state also that in most cases best accuracy rates have been obtained using dimensionality reduction techniques. Which in other words means that the selection of attributes is preferable to the raw data.

6 Conclusions and Future Work

In this paper we shown the foundations of the automated analysis of Basque impromptu greeting verses. The study proposes novel features of greeting-verses and analyses the suitability of those features in the task of automated feature classification. It is important to note that our primary goals were to establish the characteristics for the correct classification of the verses, and so to analyse their narrative structure. And, secondly, to validate different methods for categorizing Basque greeting verses.

Towards this end, we introduced different features related to improvised greeting verses and categorized them into six groups of Message, Location, Public, Event, Oneself and Miscellaneous. Then, we implemented six different approaches combining dimensionality reduction techniques and ML algorithms. One for each considered categories.

In our opinion, the most relevant conclusion is that k-nn and SMO have shown to be the most suitable algorithms for our classification task, and also, that in most cases attribute selection techniques help to improve their performance.

As a future work, we would like to assess the problem as a multi-labelling task (Zelaia et al., 2011), and see if that improves the results.

Finally, we must say that there is still much work to do in order to properly extract discourse-patterns from Basque greeting verses. To this end, we intend to use our classifiers to label larger corpora and find regular discourse patterns in them.

7 Acknowledgements

The authors gratefully acknowledge *Bertsozale Elkartea*[5] (Association of the Friends of Bertsolaritza), whose verse corpora has been used to test and develop the *Bertsobot* project.

This work has been partially conducted under the SAIOTEK projects XIMUR and POETAUTO, the Basque Government Research Team grant and the University of the Basque Country UPV/EHU, under grant UFI11/45 (BAILab).

References

Iñaki Alegria, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203.

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Harvard University Press.

Ana Cardoso-Cachopo and Arlindo Oliveira. 2003. An empirical comparison of text categorization methods. In

[5]http://www.bertsozale.com/en

String Processing and Information Retrieval, pages 183–196. Springer.

Belur V Dasarathy. 1991. Nearest neighbor ({NN}) norms:{NN} pattern classification techniques.

Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Alexis Diaz Pimienta. 2001. *Teoría de la improvisación: primeras páginas para el estudio del repentismo*. Ediciones Unión.

Andoni Egaña, Alfonso Sastre, Arantza Mariskal, Alexis Diaz Pimienta, and Guillermo Velazquez. 2004. *Ahozko inprobisazioa munduan topaketak: Encuentro sobre la improvisación oral en el mundo : (Donostia, 2003-11-3/8).* Euskal Herriko Bertsozale Elkartea.

Nerea Ezeiza, Iñaki Alegria, José María Arriola, Rubén Urizar, and Itziar Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 380–384. Association for Computational Linguistics.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305.

John J Gumperz. 1982. Discourse strategies: Studies in interactional sociolinguistics. *Cambridge University, Cambridge*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142.

Sang-Bum Kim, Hae-Chang Rim, Dongsuk Yook, and Heui-Seok Lim. 2002. Effective methods for improving naive bayes text classifiers. *PRICAI 2002: Trends in Artificial Intelligence*, pages 479–484.

Edda Leopold and Jörg Kindermann. 2002. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1):423–444.

Marvin Minsky. 1961. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30.

W Rhys Roberts and WD Ross. 2010. *Rhetoric*. Cosimo Classics.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March.

María Victoria Escandell Vidal. 2004. Aportaciones de la pragmática. *Vademécum para la formación de profesores. Enseñar español como segunda lengua (12) 1 lengua extranjera (LE)*, pages 179–197.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52.

Ana Zelaia, Iñaki Alegria, Olatz Arregi, and Basilio Sierra. 2005. Analyzing the effect of dimensionality reduction in document categorization for basque. *Archives of Control Sciences*, 600:202.

Ana Zelaia, Iñaki Alegria, Olatz Arregi, and Basilio Sierra. 2011. A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension. *Applied Soft Computing*, 11(8):4981–4990.

Iraide Zipitria, Basilio Sierra, Ana Arruarte, and Jon A Elorriaga. 2012. Cohesion grading decisions in a summary evaluation environment: A machine learning approach.