Universidad del País Vasco
Euskal Herriko Unibertsitatea

Unibertsitate Masterra
Konputazio **I**ngeniaritza eta **S**istema **A**dimentsuak

Konputazio Zientziak eta Adimen Artifiziala Saila –
Departamento de Ciencias de la Computación e Inteligencia Artificial

K I S A
I C S I

Master Tesia

Testuen irakurgarritasuna neurtzeko sailkatzaile automatikoa

Ion Madrazo Azpiazu

Tutoreak

Iñaki Inza
Konputazio Zientziak eta Adimen Artifiziala saila
Informatika Fakultatea

Montse Maritxalar
Konputazio Zientziak eta Adimen Artifiziala saila
Informatika Fakultatea

informatika fakultatea
facultad de informática

KZAA /CCIA

2014ko urriak 16

Esker onak

Eskerrik asko, **Iñaki Inza** eta **Montse Maritxalar**, proiektu honetako zuzendari lanetan bikain aritzeagatik. Zuen babes zein aholkurik gabe ezin izango litzateke proiektu hau aurrera atera.

Eskerrik asko, **Itziar Gonzales**, esperimentuetarako corpora bilatzerakoan emandako laguntzagatik.

Eskerrik asko, **Itziar Aldabe**, Laten Semantic Analysis tekniken inguruan laguntza emateagatik.

Eskerrik asko, **Ixa ikerketa taldeko kide guztiei**, urteetan sortu dituzuen erreminta guztiengatik izan ez balitz proiektu hau ezinezkoa izango bailitzateke gaur egun.

Eskerrik asko, nire **familia zein hurbileko pertsonoi**, emandako animo zein babesagatik.

Eskerrik asko, zeharka proiektu honetan parte hartu duten beste pertsona guztioi.

II

Gaien Aurkibidea

Irudien Zerrenda

Taulen Zerrenda

Laburpena

Jarraian aurkezten den dokumentuan irakaskuntza arloan laguntzeko sistema baten azterketa eta garapena azaltzen dira. Zehatzago hitz eginez, testu baten konplexutasuna neurtzeko gai den sailkatzaile baten, analisi, diseinu, garapen eta ebaluazioak aurkezten dira. Sistema garatzeko hizkuntzaren prozesamendurako teknika ezberdinak erabiltzen dituzten 6300 ezaugarri inplementatu dira. Ezaugarri hauen arteko konparaketa eta aukeraketa bat burutu da sistema optimoa lortzeko asmotan. Sistemarako algoritmo hoberena aukeratu ondoren honen emaitzak egokitzeko estrategia ezberdinak aplikatu zaizkio. Bukaeran sortzen den sistemak, literaturan aurki daitezkeen beste sistemen pareko datuak aurkezten ditu.

Sarrera

Master tesi honen helburu nagusia, testu bat irakurtzeko behar den hizkuntza maila automatikoki sailkatzeko gai den sistema baten sorkuntza da. Sistemaren helburua testu bat jaso eta hau marko europarreko mailaketaren (A1, A2, B1, B2, C1, C2) arabera sailkatzea izango da. Horretarako sailkapen automatikorako teknika klasikoez baliatzeaz gain zenbait estrategia ez hain komun erabiliko dira emaitzak problematikara egokitzen saiatzeko. Hau guztia hizkuntzaren prozesamendurako teknika ugariren laguntzari esker burutua izango da.

Arloko literatura oso zabala ez den arren, existitzen dira garatuko den sistemaren antzekoak. Adibide moduan, ebaluazioan erreferentzia gisa erabiliko den [11] sistema, frantseserako garaturiko sistema bat da. Artikulu horretan maila ezberdinetako ezaugarri linguistikoen kontaketa hutsak erabiliz europar markoko 6 mailaketen artean sailkatzeko gai den sailkatzaile bat erakusten da.

Beste hizkuntzetarako sistemak ugariagoak diren arren, euskararako ErreXail [13] izeneko sistema existitzen da soilik. Sistema hau testuak bi kategoriatan soilik sailkatzeko gai da, testu errazak eta testu zailak. Testu erraztzat umeei orientaturiko testuak jotzen dira sistema honetan eta zailtzat testu espezializatuagoak.

Dagoeneko euskararako eredu bat existitzen denez, gure sistema ez da hutsetik hasiko. Sistema garatzeko ErreXail sisteman erabiltzen diren ezaugarriak gure sisteman inplementatuko dira eta hauei ezaugarri berriak gehituko zaizkie. Gainera, gure sistemak ErreXailek baino problematika zabalagoa hartuko du kontuan, mailaketa zabalagoa erabiliko baitu.

Dokumentua honela banatzen da:

Lehen kapituluan sistema sortzeko erabili diren hizkuntzaren prozesamendurako tekniken azalpenak aurkezten dira. Bertan, testu lau bat prozesatzeko burutu behar diren pausoen eta honetarako erabiltzen diren erreminten ikuspegi orokor bat emateaz gain arloko beste zenbait tresna eta formatu espezifikoago batzuen berri ere ematen da.

Bigarren kapituluan, esperimentuan erabiliko den datu multzoari buruzko informazioa aurkezten da. Bertan, beharrezko dokumentuak nondik lortu diren azaltzeaz gain, jasan duten prestakuntzaren berri ematen da.

Hirugarren kapituluan sistemaren eraikuntzarako lehen fasea azaltzen da. Bertan sailkatzailearen ezaugarriak nola sortzen diren azaltzen da. Ezaugarri hauen sorrera, kasu askotan, lehen kapituluan aipatu diren tekniken konposaketan oinarritzen da.

Laugarren kapituluan sistemaren sorkuntzaren bigarren fasea aurkezten da. Bertan, aurreko kapituluan sortu diren ezaugarrien artean konparaketa zein aukeraketak egiteko burutu diren teknikak aurkezteaz gain, sailkapena burutzeko orduan datuak problematikara egokitzen saiatzeko zenbait estrategia ere azaltzen dira.

Bosgarren kapituluan, sistemaren ebaluazioa burutzen da. Ezaugarri konparaketak sorturiko emaitzak aztertzeaz gain sistemak sailkatzaile ezberdinekin nolako emaitzak ematen dituen ikusten da. Erabilitako zenbait egokitze estrategiaren ondorioak ere erakusten eta aztertzen dira.

Amaitzeko, seigarren kapituluan proiektuaren ondorio eta etorkizunerako lanaren berri emango da.

1 Kapitulua

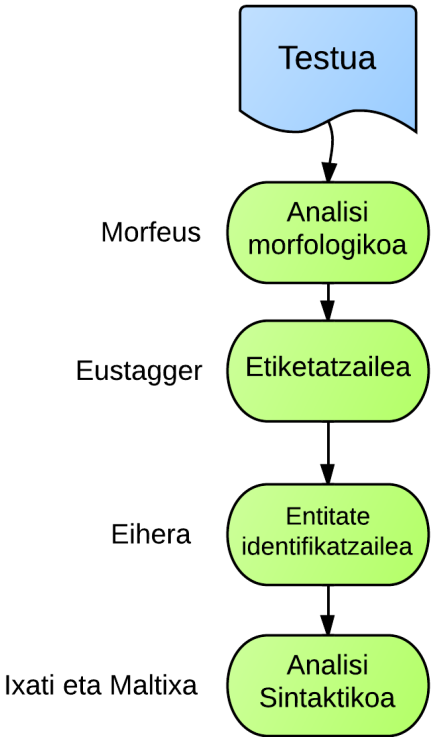
Hizkuntzaren prozesamendurako teknika eta baliabideak

Gaien Aurkibidea

Kapitulu honetan proiektua garatzeko erabili diren hizkuntzaren prozesamendurako zein bestelako tresnak aurkeztuko dira. Hala nola, erabilitako aplikazioak, formatuak eta baliabideak.

1.1 Hizkuntzaren prozesamendua

Atal honetan hizkuntzaren prozesamendurako erabilitako aplikazioak azalduko dira. Lehenik testutik hurbilen jarduten duten aplikazioei buruz hitz egingo da, pixkanaka hauek konbinatuz analizatzaile sintaktikoak azaltzera iritsi arte. 1.1 irudian ikusiko diren programen laburpen bat erakusten da. Bestalde, erabili diren zenbait baliabide zein formaturen inguruko azalpenak ere emateaz gain, Latent semantic analysis teknikaren inguruko informazioa ere ematen da.



Irudia 1.1: Analizatzaileen laburpena

1.1.1 Analisi morfologikoa

Analizatzaile morfologikoa[1] azalduko diren aplikazioetatik testutik hurbilen lan egiten duena da. Honek testuko hitzak banan-banan hartu eta morfologikoki hitz horri dagozkion aukera guztiak itzultzen ditu. Honetarako egoera finituzko transduktore bat erabiltzen du. Bere emaitza hitz bakoitzaren morfema zerrenda bat da, hau da, lema, pluraltasun, mugatasun eta bestelako markak. Honi dagokion beste zenbait informazio ere gehitzen dio, hala nola, sintagmak izan dezakeen kasua eta esaldian izan dezakeen kategoria(aditza, izena, adjektiboa...). Hona hemen adibide bat:[2]

Txoriak	**hori**	**dakar**
txori+ak(IZEARR+ABS)	horitu+0 (ADISIN+AMM)	dakar (ADT)
txori+ak(IZEARR+ERG)	hori (ADJARR)	
	hori+0 (DETERK+ABS)	

Taula 1.1: Analisi morfologikoaren adibide bat

Adibidean *Txoriak hori dakar.* esaldiaren analisi morfologikoa erakusten da. Bertan *txoriak* hitzak bi analisi ezberdin izan ditzakeela ikus daiteke. Bietan bere kategoria izen arrunta izango da, baina batean kasua absolutiboa izango da eta bestean ergatiboa. *hori* hitzak 3 aukera aurkezten ditu, aditza, adjektiboa edo determinatzailea izatekoak. Azkenik *dakar* hitzak aditz trinkoa izateko aukera soila du.

Hitzen ezaugarri guzti hauek oso garrantzitsua izango dira proiekturako. Izan ere, ezaugarri hauetariko asko sailkatzailearen sarrera ezaugarri bihurtuko dira zuzenean eta beste askok ezaugarri konplexuagoak eraikitzeko balioko dute.

1.1.2 Etiketzailea

Etiketatzailetzat (*Part of Speech tagger*) *Eustagger* aplikazioa erabili da, honek 1.1.1 atalean azaltzen den *morfeus* analizatzaile morfologikoa du integratuta. *Eustagger*en helburua analizatzaile morfologikoak ematen dituen analisi posible guztien artean, esaldi osoa kontuan hartuta, probabilitate handiena

[1] Analisi morfologikoa burutzeko Ixa ikerketa taldeko [3] *morfeus* aplikazioa erabili da.
[2] Informazio morfologikoa laburtua azaltzen da, hau garbiago egiteko.

duena aukeratzea da. Hau horrela, 1.1 ataleko analisia honela murriztuko luke:

Txoriak	**hori**	**dakar**
txori+ak(IZEARR+ERG)	hori+0 (DETERK+ABS)	dakar (ADT)

Taula 1.2: Analisi morfologikoa, etiketatzaile ondoren

1.1.3 Entitate izendatu identifikatzailea

Entitate izendatu identifikatzaile (*Named Entity Recognizer*)[3] batek testu bati buruzko informazio semantikoa ematen digu. Bere helburua hitz edo hitz multzo batek pertsona, leku edo instituzio bat errepresentatzen duen adieraztea da. Adibidez *Etxe Zuria* entitate izendatu bezala etiketatuko du Ameriketako Estatu Batuetako presidentearen etxeari buruz ari bagara, baina ez edozein etxe zuri arrunti buruz ari bagara. Hau detektatzeko datubasean gordetako entitateak zein datu meatzaritza tekniketatik lorturikoak erabiltzen ditu.

Aplikazio honen detektatutako ezaugarriak sailkatzailearen sarrera zuzen bihurtuko dira.

1.1.4 Analizatzaile sintaktikoak

Analizatzaile sintaktiko baten helburua esaldi baten sintagma zein berauen arteko dependentziak detektatzea da. Hau egiteko bi analizatzaile ezberdin erabili dira. Bat azaleko sintaxia (Ixati) burutzeko eta bestea sintaxi sakonago (Maltixa) bat burutzeko. Bi analizatzaileok aurretik aipaturiko erreminta guztiak dituzte integratuta, hau da, esaldi baten analisi sintaktikoaz gain, bere informazio morfologiko zein entitate izendatuei dagokiona ere aurkezten digute.

1.1.4.1 Sintaxi partziala

Sintaxi partzialaren[6] helburua ez da analisi guztiz zehatz bat sortzea. Bere helburua esaldi bateko *chunk*ak detektatzea da. *Chunk* bat Abbeyren definizioaren arabera, bata bestearen alboan dauden eta elkarrekin erlazioa duten

[3]Proiektuan erabili den entitate izendatu identifikatzaileak *Ehiera* du izena eta Ixa ikerketa taldean garatutako da.

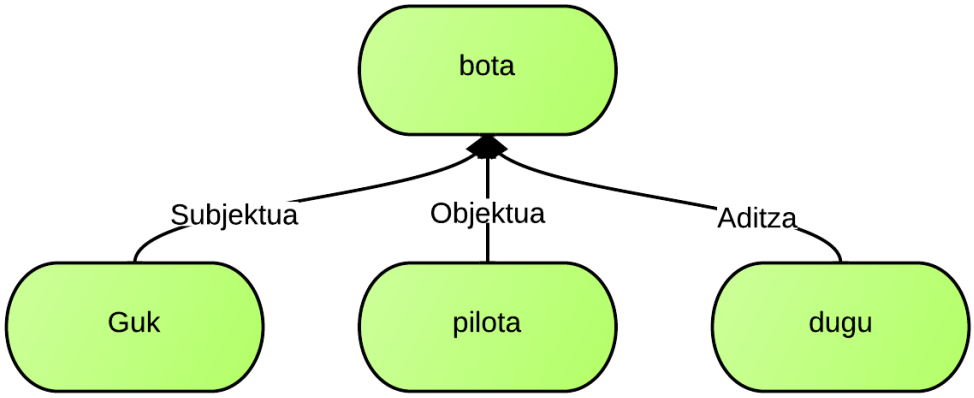
hitz multzo bat da. Honek sintagma bat osa dezake edo ez. Hona hemen adibide bat:

[Hartutako neurriek] [emaitza onak] [eman dituzte].

Izen sintagmak berdez, eta aditz sintagmak gorriz. Proiektuan azaleko sintaxia burutzeko Ixa ikerketa taldeko *Ixati*[6] aplikazioa erabili da. Bere irteera 1.2.2 atalean azaltzen den *Kyoto Annotation Framework(KAF)* formatuan eta 1.2.1 atalean ikus daitekeen formatuan erakus daiteke.

1.1.4.2 Sintaxi osoa

Sintaxi osoko analizatzaileen helburua esaldi baten zuhaitz itxurako errepresentazio bat sortzea da. Zuhaitz egitura horretan nodoak hitzak dira eta ertzak sintagmen arteko dependentziak. Sakoneko analisia burutzeko Ixa ikerketa taldeko *Maltixa* [16] aplikazioa erabili da. Bere irteera 1.2.3 atalean azaltzen den *Computational Natural Language Learning(CoNLL)* formatuan adierazten da. Hona hemen adibide bat:



Irudia 1.2: Dependentziak

Dependentzia hauek oso interesgarriak dira esaldiaren egiturari buruz informazioa ematen baitute, eta esaldien egituraketa testuen konplexutasunaren markatzaile egokitzat jo izan ohi da.

1.1.5 Latent Semantic Analysis (LSA)

Latent Semantic Analysis[9][17] dokumentuak zein terminoak semantikoki tratatzeko teknika bat da, hitzen arteko antzekotasun semantikoa kalkula-

tuz. LSA-ak bi termino semantikoki antzekotzak jotzen ditu batera agertzeko joera badute. Zenbat eta dokumentu gehiagotan batera agertu, orduan eta antzekoagoak izango dira bi termino LSA-rentzat.

Antzekotasun hauek neurtzeko agerpen matrize bat erabiltzen da. 1.3 taulan matrize honen adibide bat ikus daiteke.

	d1	d2	d3	d4	d5	d6	d7
h1	0	**1**	0	0	0	**1**	0
h2	0	0	0	1	0	0	0
h3	0	1	1	0	0	0	0
h4	0	0	0	0	0	0	0
h5	0	**1**	0	0	0	**1**	0
h6	0	1	0	0	0	0	0
h7	0	0	1	0	0	0	0
h8	0	0	0	1	0	0	1

Taula 1.3: LSA matrizearen adibide bat

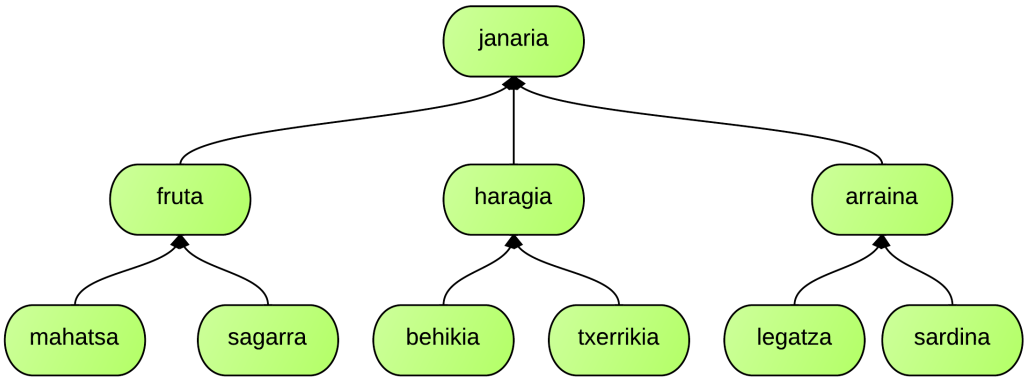
Matrizean zutabeek dokumentuak errepresentatzen dituzte eta errenkadek hitzak. Adibide moduan h1 eta h2 hitzen arteko antzekotasun semantikoa nulua izango litzateke ez baitira dokumentu berdinean batera inoiz agertzen. Aldiz, h1 eta h5 hitzen arteko antzekotasuna maximoa litzateke, h1 agertzen den guztietan h5 agertzen baita eta alderantziz.

Antzekotasuna neurtzeko neurri ezberdinak erabili ohi dira, erabiliena bi bektoreen arteko kosinua da. Hau 0 eta 1 artean normalizatua aurkitzen da eta sinplea da bektore moduan konputatzeko.

1.1.6 WordNet

WordNet [15] hizkuntza ugaritan aurki daitekeen datu-base lexikal bat da. Bertan izenak, adjektiboak, aditzak eta abar sinonimia, hiperonimia, hiponimia eta beste zenbait erlazioen arabera antolatuta aurkitzen dira. 1.3 irudian bere egituraren adibide bat ikus daiteke. Bertan nodo gorena *janari* kontzeptua da, bere hiponimoak *fruta*, *haragi* eta *arrain* direlarik. Hauek, aldi berean, bere hiponimoak dituzte. Horrela behera eta goraka jarraituz grafo erraldoi bat dago sortuta *WordNet* datu-basean.

Ohikoa da maila altuko testuetan sinonimoen erabilera aberatsago bat



Irudia 1.3: Wordneten egituraren adibide bat

aurkitzea. Erabilpen aberats hau neurtzeko gai den ezaugarriak garatzeko erabiliko da baliabide hau sisteman.

1.2 Formatuak

Jarrian proiektuan erabili diren hiru formaturen berri ematen da. Lehenik *Zatiak* formatuari buruz hitz egingo da, ondoren, *Kyoto Annotation Framework (KAF)* formatuari buruz eta amaitzeko, *Computational Natural Language Learning(CoNLL)* formatuaren berri emango da.

1.2.1 Ixatiren irteera formatua

Jarraian erakusten den formatua IXA ikerketa taldeko aplikazio askok barne formatu moduan erabiltzen duten formatu bat da. Izan ere, ondoren erakutsiko diren bi formatuak sortzeko lehenik formatu hau erabiltzen da tartekaki moduan.

Formatu honek testu baten analisi linguistikoa errepresentatzeko balio du, maila morfologiko zein sintaktikoan. Formatua hau, lerroka ordenatuta aurkitzen da. Lerro bakoitzean morfema edo lemma bat aurki dezakegu eta honen jarraian lemma edo morfema honen ezaugarri zerrenda bat hutsunez banatua. Jarraian *Jaiotza tasa baxuak eragindako arazoa izango da.* esaldiaren analisia ikus daiteke zatiak formatuan.

```
1
2 "<Jaiotza>"<HAS_MAI>" S:130/0
3       "jaiotza" IZE ARR BIZ- ZERO AORG HAS_MAI
```

```
          w23,L-A-IZE-ARR-272,lsfi32 @KM> %SIH S:130
          &ESALDI_HAS_1
"<tasa>"
       "tasa" IZE ARR ZERO AORG
          w24,L-A-IZE-ARR-275,lsfi33 @KM>
"<baxuak>"
       "baxu" ADJ ARR IZAUR- ABS NUMP MUGM
          w25,L-A-ADJ-ARR-52,lsfi36 @SUBJ %SIB
"<eragindako>" S:509/0
       "eragin" ADI SIN PART GEL ZERO
          w26,L-A-ADI-SIN-110,lsfi38
          @-JADNAG_MP_IZLG> %ADIKAT S:509 }MUGA
"<arazoa>"
       "arazo" IZE ARR BIZ- ABS NUMS MUGM
          w27,L-A-IZE-ARR-278,lsfi40 @PRED %SINT
"<izango>"
       "izan" ADI SIN PART GERO NOTDEK
          w28,L-A-ADI-SIN-107,lsfi42 @-JADNAG
          %ADIKATHAS
"<da>"
       "izan" ADL A1 NOR NR_HURA
          w29,L-A-ADL-51,lsfi44 @+JADLAG %ADIKATBU
"<$.>"<PUNT_PUNT>" S:123/0 S:148/0
      PUNT_PUNT S:123 &ESALDI_BUK_1 S:148 }MUGA
```

Bertan ikus daitekeen moduan morfemak <> ikurren artean agertzen dira errepresentatuak eta kasu gutxitan dute ondoren informazioa. Informazio gehiena lemaren ondoren agertzen da. Adibide moduan *ADI* etiketak aditza errepresentatzen du, *IZE* etiketak izena eta *BIZ-* etiketak bizigabea.

1.2.2 Kyoto Annotation Framework (KAF)

Kyoto Annotation Framework (KAF) [5] testuak linguistikoki etiketatzeko formatu bat da. *XML*n oinarritua dago eta bertan testu baten analisia gordetzen da, maila morfologiko, sintaktiko eta semantikoan. Aurretik aipatu den formatua bezala formatu hau 1.1.4.1 atalean azaltzen den Ixati analizatzailearen irteera formatu moduan erabiltzen da. Honako xehetasunak aurkezten ditu: *XML* formatua jarraitzen duenez lehenik *XML* goiburukoa dator, bertan *XML* bertsioa eta kodeketa mota zehazten direlarik. Honen jarraian dokumentuaren erro nodoa irekitzen da, *KAF* bezala izendatua.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<KAF xml:lang="eu">
```

Lehen datu bezala testuko *wordForm*ak datoz, hau da, hitzak testuan agertzen diren moduan. Hauek *wf* etiketarekin zehazten dira. *WordForm* bakoitzean berau zenbatgarren esalditik lortutako den adierazten da *sent* atributuaren bitartez.

```
<text>
<wf wid="w1" sent="1">Gaur</wf>
<wf wid="w2" sent="1">eguraldi</wf>
<wf wid="w3" sent="1">ona</wf>
<wf wid="w4" sent="1">dago</wf>
<wf wid="w5" sent="1">.</wf>
</text>
```

Hurrengo nodo mota *term* motakoa da. Honek termino bat sinbolizatzen duelarik. Termino bat hitz bat bakarrik edo hitz anitzeko unitate bat izan daiteke, adibidez *Estatu Batuak* hitz bat baino gehiagoko termino bat izango litzateke. Termino bakoitzean aurkezten den informazioari dagokionez bere lema eta eta kategoria (*pos*) agertzen dira. Lema hitz baten forma normala da, hau da, hiztegi batean aurki dezakegun forma. Kategoria, aldiz, hitzak esaldian hartzen duen partea da, hala nola, aditza, izena, adjektiboa... Termino batzuetan dagokion sintagmaren kasua ere adierazten da *case* atributuaren bidez.

```
<terms>
<!-- Gaur -->
<term tid="t1" type="open" lemma="gaur"
   pos="A.ADB-ARR">
<span>
<target id="w1"/>
</span>
</term>
<!-- eguraldi -->
<term tid="t2" type="open" lemma="eguraldi"
   pos="N.IZE-ARR">
<span>
<target id="w2"/>
</span>
</term>
<!-- ona -->
```

```
<term tid="t3" type="open" lemma="on" pos="G.ADJ-ARR"
   case="ABS">
<span>
<target id="w3"/>
</span>
</term>
<!-- dago -->
<term tid="t4" type="open" lemma="egon" pos="V.ADT">
<span>
<target id="w4"/>
</span>
</term>
</terms>
```

Azken nodo mota bezala *chunk*-ak ditugu. *Chunk* bat termino multzo bat da, esaldian funtzio sintaktiko bat izan dezakeena. *Chunk* analisia nolabaiteko analisi sintaktiko partzial bat da. *Chunk* bakoitzean aditz edo izen sintagma bat den erakusten zaigu *phrase* atributuaren bidez. Hemen ere chunk batzuetan bere kasua erakusten zaigu, adibidez *eguraldi ona* chunka absolutibo kasukoa dela adierazten da.

```

<chunks>
<!-- Gaur -->
<chunk cid="c1" head="t1" phrase="NP">
<span>
<target id="t1"/>
</span>
</chunk>
<!-- eguraldi ona -->
<chunk cid="c2" head="t2" phrase="NP" case="ABS">
<span>
<target id="t2"/>
<target id="t3"/>
</span>
</chunk>
<!-- dago -->
<chunk cid="c3" head="t4" phrase="VP">
<span>
<target id="t4"/>
</span>
```

```
21 </chunk>
22 </chunks>
```

1.2.3 Computational Natural Language Learning(CoNLL)

Computational Natural Language Learning(CoNLL) formatua 1.1.4.2 atalean azaltzen den *Maltixa* analizatzaile sintaktikoaren irteera moduan erabiltzen den formatu estandar bat da. Formatu hau taula formatu bat da, tabulatzailez banaturiko zutabez osatua eta lerro saltoz banaturiko errenkadez. Errenkada bakoitzean termino baten inguruko informazioa erakusten da. Taulak zuhaitz egitura bat errepresentatzen du, egitura hau *id* eta *burua* atributuen bitartez lortzen da. Maltixaren irteera baten adibidea 1.4 taulan eta 1.2 irudian ikus daiteke, taula moduan zein zuhaitz errepresentazioan hurrenez hurren.

id	Hitza	Burua	Dep.	Lema	Ezaug.	Kat.	Azpikat.
1	Guk	3	ncsubj	gu	KAS:ERG	IOR	PERARR
2	pilota	3	ncobj	pilota	KAS:ABS	IZE	ARR
3	bota	0	ROOT	bota	ADM:PART	ADI	SIN
4	dugu	3	auxmod	edun	_	ADL	ADL
5	.	4	PUNC	.	_	PUNT	PUNT

Taula 1.4: Maltixaren irteera

Honako zutabeak aurki ditzakegu bertan:

- **Id**: Hitzaren identifikatzailea
- **Hitza**: Hitza esaldian zetorren bezala.
- **Burua**: Hitzaren burua, zuhaitz egituran gurasoa izango litzatekeena.
- **Dependentzia**: Uneko hitzak bere buruarekiko (gurasoa) duen dependentzia sintaktikoa, subjektua, objektua...
- **Lema**: Hitzaren lema erakusten du, hau da, hitza hiztegian agertzen den forma normalean.
- **Ezaugarriak**: Bertan ezaugarri ezberdinak aurki ditzakegu, kasua, numeroa, aspektua...

- **Kategoria**: Hitzaren kategoria adierazten du, aditza, izena, adjektiboa...
- **Azpikategoria**: Kategoria atributuaren zehaztapen bat da, izenen kasuan arrunta edo berezia ote den adierazten da bertan adibidez. Entitate izendatuen kasuan ere hemen adierazten da informazio hau.

2 Kapitulua

Corpus azterketa

Gaien Aurkibidea

Ikerketa lanarekin aurrera jarraitu ahal izateko, bai probak burutzeko zein modulu batzuk entrenatzeko, corpusak lortzeko beharra ikusi da. Jarraian, beharrezko corpusak identifikatu, lortu eta bakoitza prestatzeko burutu diren lanak aurkezten dira.

2.1 Beharren identifikazioa

Tesiaren helburua, aurretik aurkeztu den legez, testuen irakurgarritasun maila modu automatikoan neurtzeko teknikak garatzea da. Beraz, irakurgarritasun maila ezberdinetako testuak lortzea funtsezkoa izango da tesiaren garapenerako, probak egin, entrenatu eta berau ebaluatzeko.

Gainera, testuen maila neurtzeko hizkuntza marko europarra hartu denez estandartzat, mailaketa honetan aurkitzen den corpora lortzea beharrezkoa izango da. Edo azken kasuan, marko europarrarekiko berdintasunak egitea posible litekeen corpora bat.

Bestetik, maila bakoitzean 300 testu inguru edukitzea beharrezkotzat jo da, testu multzo adierazgarri bat izateko, testu bakoitzak gutxienez 300 hitz inguru izango dituelarik.

Azkenik, proiektuan erabiliko diren hizkuntzaren prozesamendurako tresna batzuen beharrak betetzeko domeinu orokorreko corpus baten beharra ere identifikatu da.

2.2 Analisia

Ixa ikerketa taldean[3] existitzen diren corpusak identifikatzeaz gain sarean aurki daitezkeen baliabideak ere analizatu dira. Honakoak izan dira identifikaturiko baliabide posibleak:

2.2.1 Egunkaria corpusa

Egunkaria corpusa *Egunkaria* egunkariak urteetan zehar argitaratutako testuen bilduma bat da. Eguneroko berrien inguruko informazioa argitaratu izanak domeinu orokorreko corpus moduan erabiltzeko aproposa bihurtzen du, bertan arlo askotariko testuak aurki baitaitezke. Corpus hau Ixa ikerketa taldeko zerbitzarietatik hartu da.

2.2.2 Wikipedia

Wikipedia[1] erlazionaturiko dokumentu multzo handi batez osatua dago. Erlazio hauek erabiliz domeinu askotariko testuak lortzea posible da, horrela domeinu orokorreko zein domeinu espezifikoko bilduma bat sortzea ahalbidetuz.

2.2.3 Ikasbil

Ikasbil[2] euskara ikasten edo irakasten ari direnentzat orientaturiko web atari bat da. Bertan euskara ikasteko modu askotariko baliabideak aurki daitezke, bideoak, entzutezko ariketak, irakurmen ariketak eta abar. Webgune hau bereziki interesgarria egiten duena bere mailaketa da, bertako material guztiak hizkuntza marko europarraren arabera mailakatuak aurki baitaitezke. 2.1 taulan maila bakoitzean aurki daitekeen testu kopurua erakusten da

A1	A2	B1	B2	C1	C2
1	24	481	2497	748	39

Taula 2.1: Ikasbilen aurki daitezkeen testu kopuruak mailaka

2.2.4 Elhuyar

Elhuyar web atarian erreportaia eta albiste espezializatu ugari aurki daitezke. Ixa ikerketa taldean bertatik erauzitako corpus bat existitzen da 100 albiste eta 100 erreportaiaz osatua. Testuok espreski mailakatuak ez dauden arren, taldeko hizkuntzalari batek C2 mailakotzat jo daitezkeela baieztatu du, hauen espezializazio zein hizkuntza maila altuagatik.

2.2.5 Administrazio corpusa

EHUko Euskara Zerbitzuak 6 liburu oso itzuli zituen gazteleratik, hauekin itzulpengintza automatikorako corpora bat sortzeko asmotan. Liburu hauen artean interesgarritzat jo den liburua erakundeen administrazioari buruz hitz egiten duen liburua da, bere teknikotasun zein idazkera maila altuagatik. Liburu hau 102 testu txikiagotan zatikatua aurkitzen da eta hau ere C2 mailako corpustzat sailkatu zuen hizkuntzalari batek.

2.3 Ondorioak eta aukeraketa

Baliagarriak gerta daitezkeen corpusak identifikatu ondoren garbi gelditu da hizkuntzaren arabera mailakaturik aurkitu daitekeen testu kopuru ez dela oso zabala. Espreski mailakaturik aurkitzen den corpus bakarra identifikatu da, ikasbilekoa hain zuzen. Gainera, corpus honetako maila bakoitzeko testu kopurua oso desorekatua da, maila batzuetan testu gutxiegi izateko punturaino. A1, A2 eta C2 mailetan dagoen testu kopurua hasieran ezarritako 300 testuko minimotik urrun aurkitzen da, 1, 24 eta 39 hurrenez hurren. C2 mailarako bestelako baliabide batzuk aurkitu direnez maila hau osatzea posible izango da, Elhuyarreko erreportaia eta albisteak eta administrazioko testuak gehituz 341 testu lortu baitira guztira.

Baliabide mugak ikusirik, nahiz eta hasiera batean marko europarreko maila guztiekin lan egin nahi izan, mailaketak murriztea erabaki da. Honela, tesian zehar esperimentu guztiak 4 mailarekin burutuak izan dira, B1, B2, C1 eta C2 mailak zehazki. Saiakera batzuk egin ziren arren, ezinezkoa izan zen beheko bi mailetarako corpora egokia aurkitzea.

Amaitzeko, modulu batzuetarako beharrezkoa izango den corpus orokortzat egunkaria corpusa erabiltzea erabaki da. Honen arrazoia gradu amaierako proiektuan [8] wikipediatik corpusak erauztean izandako esperientzia da. Bertatik lorturiko testuek zarata ugari izateko joera dute, figuren erreferentziak, estekak... Arazo hauek ekiditeko asmotan, wikipedia corpusa baztertu eta egunkaria corpusa erabiltzea erabaki da.

2.4 Eskurapena

Jarraian corpus bakoitza nola eskuratu den azaltzen da. Prozesu honetan ez da eskuraketa hutsa burutu, corpusak lortzeaz gain guztiak gerorako komenigarria izango den formatuan jarri dira.

2.4.1 Egunkaria corpusa

Egunkaria corpusa Ixa ikerketa taldeko zerbitzarietan aurkitzen da. Corpusa testu lauan aurkitzen da, beraz ez da formatu aldaketarik burutu behar izan. Aldiz, bere tamaina handiegia da esperimentuak modu azkar batean burutu ahal izateko. Horregatik, zati bat bakarrik hartu da. Zatia hartzean corpusaren izaera orekatua mantendu nahi izan da. Corpusa domeinu orokorrekoa da bere osotasunean, baina ez du zertan hala izan behar zati bat hartzen badugu. Adibidez, egun guztietako lehen 15 dokumentuak hartzen baditugu, domeinu orokorra izatetik, ekonomia domeinukoa izatera pasa dai-

teke, egunero lehen orrietan ekonomiaz hitz egiten baita gehien. Horregatik zatia erauzteko moduak egunak bere osotasunean errespetatu ditu. Guztira bi hilabete oso hartu dira 200.000 hitz inguru osatzeko.

Gerora begira, corpusa linguistikoki prozesatu da eta KAF (Kyoto Annotation Format) formatuan gorde da.

2.4.2 Ikasbil

Corpusa erauzteko web armiarma bat garatu da Java lengoaia eta Jsoup[4] liburutegia baliatuz. Liburutegi honek HTML dokumentuak DOM eredua jarraituz prozesatzeko erraztasunak ematen ditu, horrelako armiarma baten garapena asko erraztuz. Armiarmak ikasbil web atariko dokumentu edukiak banan-banan zeharkatzen ditu, bideo eta entzutezko edukiak baztertuz. Prozesua burutu ostean dokumentu guztiak mailaka ordenaturik uzten dira direktorio batean.

Behin testu guztiak lortuta, hauek linguistikoki prozesatuak izan dira bi tresna ezberdinekin. Batetik chunketan oinarrituriko *Zatiak*[6] analizatzailearekin prozesatu dira, esaldi mugak eta aposizioak markatzeko tresna bat ere erabiliz. Bestetik dependentzietan oinarrituriko *Maltixa*[16] analizatzailearekin. Analizatzaile hauen eta beren formatuen inguruan informazio gehiago nahi izanez gero ikus 1 kapitulua.

2.4.3 Elhuyar eta administrazio testuak

Elhuyarreko erreportai eta albisteak zein administrazio testuak Ixa ikerketa taldeko zerbitzarietan aurkitzen dira. Beren jatorrizko formatua testu laua da. Hauek ere bi analizatzailerekin prozesatu dira, chunketan oinarrituriko *Zatiak* analizatzailea eta dependentzietan oinarrituriko *Maltixa* analizatzailea.

3 Kapitulua

Konplexutasun sailkatzailea: iragarpen ezaugarriak

Gaien Aurkibidea

Sailkatzailea garatzeko bi fase garrantzitsu burutu behar izan dira. Batetik ezaugarriak definitu eta jadanik prestaturiko corpusetatik erauzi behar izan dira. Bestetik, jadanik ezaugarri guztiak lortuta izanik, hauen gainean ikasketa metodo ezberdinak aplikatu dira, ahalik eta sailkatzaile doituena lortzeko asmotan. Atal honetan lehen fasearen deskribapen bat burutzen da, sorturiko ezaugarriak banan banan deskribatuz.

Dokumentuaren hasieran aipatzen den moduan, sortu nahi den sistema ez da guztiz hutsetik hasiko. ErreXail [13] sisteman erabiltzen diren ezaugarriak oinarritzat hartuko dira eta hauei ezaugarri berri gehiago gehitu. Ezaugarri zerrenda oinarritzat erabiliko den arren hauek berriro inplementatu behar izan dira, sistema bateratuago, eta beraz, mantengarriago bat sortzeko asmotan.

3.1 ErreXaileko ezaugarriak

Lehen atal honetan ErreXail sistemak erabiltzen dituen ezaugarriak azalduko dira. Ezaugarriak artikuluan ematen den sailkapen berdina jarraituz erakutsiko dira. Ezaugarriak modu laburtu batean azalduko dira, hauek osotasunean ikusi nahi izanez gero ikus I eranskina.

3.1.1 Ezaugarri orokorrak

Ezaugarri orokorrek dokumentua bere osotasunean hartzen dute kontuan. Lau ezaugarri orokor definitu dira:

- Esaldiko hitz kopurua batazbestean.
- Esaldiko sintagma kopurua batazbestean.
- Hitzeko karaktere luzera batazbestean.
- Behin agertzen diren lemak lama guztien kopuruarekiko normalizatuta.

3.1.2 Ezaugarri lexikoak

Ezaugarri lexikoak hitzaren forma oinarrizkoaren, hau da, lemaren inguruko ezaugarriak dira. Ezaugarri mota ezberdinak definitzen dira maila honetan.

- **Kategorien maiztasunak** : izenak, aditzak, adberbioak...

- **Pertsona ezaugarrien maiztasunak** : nor, nor-nori, nor-nori-nork eta nor-nork motako aditzak.
- **Laburtzapen moten maiztasunak** : laburtzapen orokorrak, siglak edo akronimoak.
- **Bestelako ezaugarrien maiztasunak** : entitateak eta aditz modalak.

3.1.3 Ezaugarri morfologikoak

Ezaugarri morfologikoek lemek jasandako forma aldaketak aztertzen dituzte. Jarraian morfologiari dagokionez erabili diren ezaugarriak erakusten dira.

- **Kasuen maiztasunak** : akusatiboa, inesiboa, datiboa...
- **Aspektu marken maiztasunak** : burutua, ez burutua...
- **Aditz aldi marken maiztasunak** : lehenaldia, geroaldia...
- **Aditz modu marken maiztasunak** : indikatiboa, subjuntiboa...
- **Elipsi marken maiztasunak** : izen elipsiak, aditz elipsiak...

3.1.4 Ezaugarri morfosintaktikoak

Maila morfosintaktikoan 5 ezaugarri definitu dira. Hauek jarraian deskribatzen dira.

1. Izen sintagma kopurua, esaldi kopuruarekiko normalizatua.
2. Izen sintagma kopurua, sintagma kopuruarekiko normalizatua.
3. Aditz sintagma kopurua, esaldi kopuruarekiko normalizatua.
4. Aditz sintagma kopurua, sintagma kopuruarekiko normalizatua.
5. Aposizio kopurua, sintagma kopuruarekiko normalizatua.

3.1.5 Ezaugarri sintaktikoak

Ezaugarri sintaktikoan esaldiaren egituraren berri ematen dituen ezaugarriak dira. Bertan menderakuntza sintagmen maiztasunak soilik erabili dira ezaugarritzat, konpletiboa, moduzkoa...

3.1.6 Ezaugarri pragmatikoak

Maila pragmatikoan esaldien arteko fenomenoak aztertzen dira. Jarraian maila honetan erabili diren ezaugarriak aurkezten dira.

- **Lokailuen maiztasunak** : emendiozkoak, aurkakoak, kausazkoak...
- **Juntagailuen maiztasunak** : emendiozkoak, hautakariak eta aurkariak.

3.2 Gehitutako ezaugarri berriak

Atal honetan ErreXaileko ezaugarriez aparte sisteman gehitu diren ezaugarri berriak deskribatzen dira.

3.2.1 Dependentzia sintaktikoak

Dependentzia sintaktikoen esaldi bateko sintagmen arteko erlazioak erakusten dituzte. Guztira 30 dependentzia sintaktiko izan dira kontuan. Haueta-riko bakoitzaren agerpen kopurua kontatzen da eta hitz kopuruarekiko normalizatu. 3.1 taulan kontuan hartu diren dependentziak erakusten dira, aurrerantzean hauek erreferentziatzeko erabiliko den etiketarekin batera.

3.2.2 Kontaketa konposatuak

Esaldi baten konplexutasuna neurtzeko batzuetan ez da nahikoa ezaugarriak banan banan neurtzea. Zenbait egitura detektatzeko esaldiko hitzak binaka edo gehiagonaka tratatzea beharrezkoa izaten da. Intuizio hau jarraian agertzen diren bi ezaugarri motekin aplikatua izan da.

Kategoriak

Kategoria batzuk beste batzuen alboan asko agertzeak egitura konplexu batzuen markatzaile izan daitezke. Adibide moduan hiru aditz kategoriako hitz bata bestearen alboan agertzea egitura konplexu baten erakusle da eta maila altuko testuetan gehiago agertzen da. Horregatik, kategoriak binaka eta hirunaka izan duten agerpen kopurua kontatu da eta hitz kopuruarekiko normalizatu. 15 kategoria ditugularik, binakako agerpenak neurtzen dituzten 225 ezaugarri berri sortu dira eta hirunakako agerpenak neurtzen dituzten 3375 ezaugarri.

Dependentzia-etiketa	**Deskripzioa**
aditz	nagusi Aditza
aponcmod	Aposizioa (ez-perpausa)
apocmod	Aposizioan dagoen mendeko perpaus jokatua
apoxmod	Aposizioan dagoen mendeko perpaus ezjokatua
arg mod	Etiketa semantikoa
auxmod	Aditz laguntzailea
ccomp obj	Mendeko perpaus osagarri jokatua, objektua
ccomp subj	Mendeko perpaus osagarri jokatua, subjektua
cmod	Mendeko perpaus jokatua; adizlaguna edo izenlaguna
detmod	Determinatzailea
entios	Entitate-osagaia
galdemod	Aditzaren indartzailea
gradmod	Graduatzailea
haos	Hitz anitzekoaren osagaia
itj out	Interjekzioa
lot	Loturazko elementuak
lot at	Lokailuak
menos	Menderagailu-osagaia
ncmod	Adizlaguna
ncmod	Modifikatzailea
ncpred	Predikatiboa (ez-perpausa)
ncobj	Objektua (ez-perpausa)
ncsubj	Subjektua (ez-perpausa)
nczobj	Zehar-objektua (ez-perpausa)
postos	Postposizio-osagaia
prtmod	Partikulak; aditzarekin agertu ohi direnak
xcomp obj	Mendeko perpaus osagarri ezjokatua, objektua
xcomp subj	Mendeko perpaus osagarri ezjokatua, subjektua
xcomp zobj	Mendeko perpaus osagarri ezjokatua, zehar-objektua
xmod	Mendeko perpaus ezjokatua; adizlaguna edo izenlaguna
xpred	Mendeko perpaus ezjokatua, predikatiboa

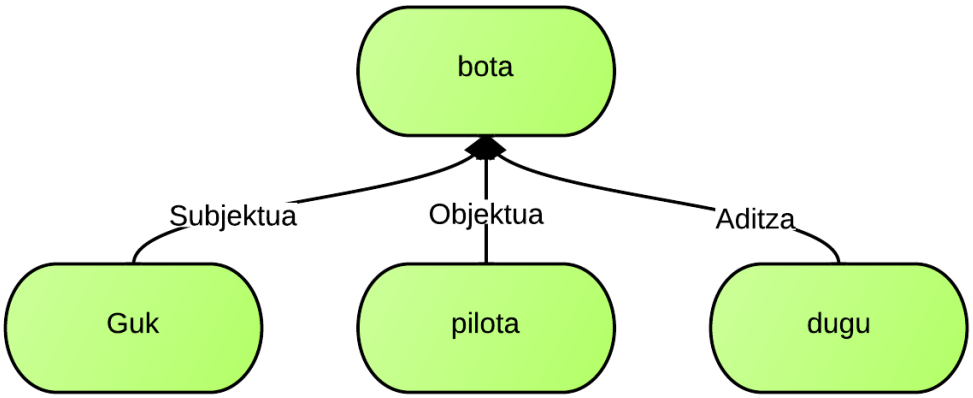
Taula 3.1: Dependentziak

Dependentzia sintaktikoak

Dependentziekin intuizio berbera jarraitu da. Esaldiaren zuhaitz egituran patroi konkretu batzuk detektatzeko asmotan dependentziak binaka kontatu dira guraso-ume erlazioak kontuan izanik. Honek 900 ezaugarri berri sortu ditu.

3.2.3 Zuhaitz sintaktikoaren sakonera

Maltixa[16] analizatzaileak esaldi baten zuhaitz errepresentazio bat ematen digu. Zuhaitz honek esaldiaren egitura sintaktikoa errepresentatzen du dependentziak eta nodoak erabiliz. 3.1 irudian *Guk pilota bota dugu* esaldiaren zuhaitz errepresentazioa ikus daiteke.



Irudia 3.1: Zuhaitz sintaktiko adibidea

Zenbat eta esaldi bat konplexuagoa izan, orduan eta konplexuagoa izango da bere sintaxia, eta ondorioz, bere zuhaitz errepresentazioa. Horregatik, zuhaitz honen sakonera neurtzea, esaldi baten konplexutasunaren markatzaile ona izan daitekeela uste da.

Ezaugarri hau osatzeko, testuko esaldi guztien zuhaitz errepresentazioen sakonerak neurtu dira eta batazbestekoa kalkulatu. Ezaugarri honi RAT_SAKONERA_SENT izena eman zaio.

$$sakonera = \sum_{i=1}^{n} \frac{sakonera(s_i)}{n} \quad (s_i = esaldia; n = esaldi\ kopurua) \quad (3.1)$$

3.2.4 Sinonimoen erabilera

Maila altuetan agertzeko joera duen beste fenomeno bat sinonimia da. Errepikapenak ekiditeko asmoz eta testuak aberasteko asmoz hitz berdinaren sinonimo ezberdinak erabiltzeko joera dago. Horregatik fenomeno hau detektatzeko gai den ezaugarri bat sortzea interesgarritzat jo zen.

Wordnet milaka hitz gordetzen dituen datu base lexikal bat da. Bertan hitzak *synset* izeneko egituretan ordenatuak daude. Synset batek esanahi bat errepresentatzen du eta bere barnean esanahi hau erreferentziatzeko balio duten sinonimo guztiak aurki ditzakegu. Gainera, synset ezberdinak elkarrekin erlazionaturik daude hiponimia hiperonimia eta beste erlazio askoren bitartez, grafo erraldoi bat sortuz. Wordneti buruz gehiago jakin nahi izanez gero ikus 1.1.6 atala.

Beraz, teknikotasunak albo batera utzita, Wordneten kontzeptuak eta hauen sinonimoak ditugu eskura. Bi unitate hauek erabiliz testuko sinonimo aberastasuna neurtzeko neurri bat proposatu da.

Honetarako lehenik, testuko hitzak erabiliz bertan existitzen diren kontzeptu guztiak erauzten dira Wordnetetik. Ondoren, kontzeptu bakoitzarentzat erreferentzia egiten dioten eta testuan agertzen diren sinonimo guztiak lortzen dira. Behin prozesu hau jarraiturik honako egitura bat lortzen da.

- Kontzeptua: k1
  - Sinonimoa : s1
  - Sinonimoa : s2
- Kontzeptua: k2
  - Sinonimoa : s3
- Kontzeptua: k3
  - Sinonimoa : s4
  - Sinonimoa : s4
  - Sinonimoa : s5
  - Sinonimoa : s5
  - Sinonimoa : s6

Hau da, k1 kontzeptua bi aldiz agertu da testuan s1 eta s2 formatan. K2 kontzeptua behin agertu da s3 forman. Amaitzeko, k3 kontzeptua 5 aldiz agertu da, bi aldiz s4 forma, bi aldiz s5 forma eta behin s6 forman.

Egitura hori lorturik, kontzeptu bakoitzaren heterogenotasuna neur dezakegu idazleak egin duen sinonimoak erabiltzeko saiakera neurtzeko. Horretarako Shannonen entropia balioa erabili daiteke. Shannonen entropia [18] informazio iturri baten ezjakintasuna neurtzeko diseinatu zen arren, multzo baten heterogenotasuna neurtzeko ere baliagarria da. Honela kalkulatzen da:

$$H(X) = \sum_i p_{x_i} \log_2(p_{x_i}) \tag{3.2}$$

X gure kasuan kontzeptua izango litzateke eta x_i sinonimo bakoitza. $P(x_i)$ X multzotik elementu bat ausaz aukeratuta x_i elementua lortzeko probabilitatea litzateke.

Estatistiko honekin balio handia emango genieke sinonimo ezberdin asko dituzten kontzeptuei eta balio txikia sinonimo ezberdin gutxiago dituzten kontzeptuei. Behin kontzeptu bakoitzaren entropia edo sinonimo aberastasuna neurturik, guztien batazbesteko bat burutzen da testuaren sinonimo aberastasuna neurtzeko.

$$aberastasuna = \sum_{i=1}^{n} \frac{aberastasuna(k_i)}{n} \quad (k_i = kontzeptua; n = kontzeptu\ kopurua) \tag{3.3}$$

3.2.5 Jarraitasun semantikoa

Jarraitasun semantiko moduan deitu dugun fenomenoa ere nahiko fenomeno komun eta intuitiboa da. Ulermen maila baxuko irakurleentzat eginiko testuak, irakurleak testua jarraitzeko arazoak izango dituela suposatuz idazten dira. Maila hauetako irakurleek testuko hitz ugari ez dituzte ulertzen eta bertako hari argumentala galtzeko erraztasun gehiago dute ulermen handiko irakurleek baino. Horregatik, testu hauetan, esalditik esaldira gertatzen diren kontzeptu aldaketak normalean leunak izaten dira. Aldiz, konplexutasun maila altuagoko testuetan aldaketak bizkor gerta daitezke testuan, irakurlea hauek jarraitzeko gai izango dela suposatzen baita.

Proposatu nahi den ezaugarri honen helburua fenomeno hau detektatzea da. Hau da, esalditik esaldira alderdi semantikotik dagoen jauzia neurtzea

nahi da.

Latent Semantic Analysis (LSA) dokumentuak semantikoki analizatzeko teknika bat da. LSA-ri buru gehiago jakiteko ikus 1.1.5 atala. LSAk bi hitzen arteko hurbiltasun semantikoa neurtzeko balioko du [1].

Behin bi hitzen arteko hurbiltasun semantikoa neurtzea posible izanda, hurrengo pausoa bi esaldiren artekoa neurtzea izango litzateke. Honen kalkulua 3.2 taulan ikus daiteke.

	h1	h2	h3	h4	max
h5	0	0.1	0.5	0	0.5
h6	0	0	0	0	0
h7	0.2	0.4	0.9	0	0.9
h8	0.3	0.4	0.9	1	1
					batazb:0.6

Taula 3.2: Jarraitasun semantikoaren kalkulua

H1, h2, h3 eta h4 hitzek lehen esaldia osatzen dute eta h5, h6, h7 eta h8 hitzek bigarren esaldia osatzen dute. Taulan hitz bakoitzaren arteko LSA balioak erakusten dira. Bigarren esaldiko hitz bakoitzerako lehengo esaldiko hitzik hurbilena zein den kalkulatzen da. Behin hitz hurbilenak zein diren kalkulaturik hauen guztien batazbestekoa lortzen da. Horrela bi esaldiren arteko hurbiltasun semantikoa lor dezakegu.

Behin esaldien arteko hurbiltasuna lorturik guztien arteko batazbesteko bat kalkulatzen da eta honek testuko jarraigarritasun semantikoa iragartzeko balio du.

[1] LSAk funtzionatzeko corpus orokor batean entrenatua izan behar du. Gure kasuan Egunkaria corpusean entrenatua izan da. Corpus honi buruz informazio gehiago nahi izanez gero ikus 2.4.1 atala

4 Kapitulua

Konplexutasun sailkatzailea: Algoritmoak

Gaien Aurkibidea

Kapitulu honetan sailkatzailearen garapenaren bigarren fasea erakusten da. Hau da, behin testuak eta ezaugarri guztiak prestaturik izanda, datuetatik ahalik eta sailkatzailerik doituena lortzen saiatuko gara. Honetarako, lehenik lortu diren ezaugarrietatik komenigarrienak aukeratzeko erabili diren metodoak azalduko dira. Amaitzeko, emaitzak hobetzen saiatzeko balioko duten meta-algoritmo batzuk erakutsiko dira.

OHARRA: Jarraian erakusten diren metodo askok aldagai diskretuekin soilik egiten dute lan. Aldiz, erabiliko diren aldagai gehienak jarraiak dira. Kasu horietan, aurkakorik esaten ez bada, metodoa aplikatu aurretik aldagaiak diskretizatuak izango dira. Diskretizazio metodotzat multzoen klase aldagaien entropia minimizatzen duen diskretizazio teknika bat erabili da, [10] artikuluan erakusten dena, hain zuzen.

4.1 Ezaugarri aukeraketa algoritmoak

Ezaugarri aukeraketa sailkapen automatikorako fase garrantzitsu bat da. Honen helburua ezaugarri zerrenda osotik komenigarrienak izan daitezkeen ezaugarriak aukeratzea da. Ezaugarri aukeraketa burutzeak kasu askotan lagun dezake. Hala nola, datu kopurua handiegia bada eta bere tamainagatik ikasketa denbora handiegiak baditu, aldagaiak murrizteak prozesu hauen abiadura hobe dezake, sailkatzailearen asmatze tasak antzeko mantenduz. Beste kasu batzuetan, zenbait ezaugarrik zarata sortu dezakete datuetan inolako informazio baliagarririk eman gabe, ezaugarri hauek kentzeak sailkatzailearen emaitza orokorrak hobe ditzake.

Gure kasura itzuliz, 6200 ezaugarri inguru ditugu bilduta. Beraz, ezaugarri kopurua nahiko handia da, dugun adibide kopurua kontutan izanda, gainera informazio gutxi emango duten ezaugarri ugari egongo diren susmoa dago. Horregatik ezaugarri aukeraketa burutzea ezinbestekotzat jo da sailkatzailea burutu aurretik. Helburua 50 ezaugarri azpitik lortzea izango delarik.

Jarraian, erabili diren ezaugarri aukeraketa metodoak aurkezten dira. Lorturiko emaitzak ikus ahal izateko ikus 5.2 atala.

4.1.1 Information gain

Information gain metodoak, a ezaugarria jakinik gero klaseak irabaziko lukeen informazioa neurtzen du. Horretarako klase ezaugarriak bere horretan

duen entropia eta ezaugarri berria gehituz sortzen diren multzoek duten entropia konparatzen dira.

Irakasgaia	**Gainditu**
Matematika	Bai
Historia	Ez
Filosofia	Bai
Matematika	Ez
Historia	Ez
Filosofia	Bai

Taula 4.1: Sailapen adibidea: Ikasleek aukeraturiko irakasgaia eta berau gainditu izana.

4.1 taulan sailkapena burutzeko adibide bat ikus daiteke. Bertan 6 ikasle daude errepresentatuta, aukeratu duten irakasgaia eta berau gainditu duten edo ez adierazten duten bi ezaugarriren bitartez. *Gainditu* iragarri beharreko klasea da eta *irakasgaia* ezaugarri bat. Bertan information gain estatistikoa aplikatuko dugu, *irakasgaia* ezaugarria erabiliz zenbat informazio irabaziko genukeen kalkulatzeko.

Lehenik klaseak bere horretan duen entropia neurtuko dugu, Shannonen entropia [18] erabiliz. X multzoa izanik eta x_i multzoko balio bakoitza, honela kalkulatzen da:

$$H(X) = \sum_i p_{x_i} \log_2(p_{x_i}) \tag{4.1}$$

Gure kasura ekarriz, honakoa litzateke *Gainditu* klasearen entropia. Hau da, *Gainditu* klaseak har ditzakeen bi balioetarako entropia edo ezjakintasun maximoa.

$$H(GAINDITU) = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1 \tag{4.2}$$

Irakasgaia ezaugarria erabiliz sailkapen taula hiru azpitauletan bana dezakegu. Azpitalde hauek sortuz:

Irakasgaia	Gainditu
Matematika	Bai
Matematika	Ez

Taula 4.2: Sailkapen adibidea: Matematika irakasgaia aukeratu duten ikasleen azpimultzoa

Lehen azpimultzoa 4.2 taulan ikus daiteke. Bertan klaseak erakusten duen entropia honakoa da:

$$H(GAINDITU|MATEMATIKA) = -[0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1 \tag{4.3}$$

Irakasgaia	Gainditu
Historia	Ez
Historia	Ez

Taula 4.3: Sailkapen adibidea: Historia irakasgaia aukeratu duten ikasleen azpimultzoa

Bigarren azpimultzoa 4.3 taulan ikus daiteke. Bertan klaseak erakusten duen entropia honakoa da:

$$H(GAINDITU|HISTORIA) = -[1 \log_2 1] = 0 \tag{4.4}$$

Irakasgaia	Gainditu
Filosofia	Bai
Filosofia	Bai

Taula 4.4: Sailkapen adibidea: Filosofia irakasgaia aukeratu duten ikasleen azpimultzoa

Hirugarren azpimultzoa 4.4 taulan ikus daiteke. Bertan klaseak erakusten duen entropia honakoa da:

$$H(GAINDITU|FILOSOFIA) = -[1 \log_2 1] = 0 \tag{4.5}$$

Hiru azpimultzoak kontuan hartuta, *Gainditu* aldagaiak *Irakasgaia* aldagaia kontuan izanik izango lukeen entropia honakoa litzateke:

$$H(GAINDITU|IRAKASGAIA) = \frac{2}{6} \times 1 + \frac{2}{6} \times 0 + \frac{2}{6} \times 0 = \frac{2}{6} = 0.33 \quad (4.6)$$

Behin bi kasuetako entropiak kalkulaturik hauek konparatzea gelditzen da irabazitako informazioa zenbatekoa izan den ikusteko:

$$IG(Y|X) = H(Y) - H(Y|X) \quad (4.7)$$

Gure kasurako:

$$IG(GAINDITU|FILOSOFIA) = 1 - 0.33 = 0.66 \quad (4.8)$$

Amaitzeko, burutu den adibidean ikus daitekeen, moduan, metodo hau metodo unibariatu bat da, hau da ez da inoiz kontuan hartzen ezaugarriek beraien artean izan dezaketen erlazio edo informazio gainezarpena. Bestalde, formuletan ikus daitekeen moduan, metodo honek ezaugarri diskretuak erabili ditzake soilik, hau da, metodoa erabiltzeko aurrediskretizazio bat erabili behar izango da.

4.1.2 Correlation feature selection

Correlation Feature Selection (CFS)[14] metodoa proiektuan erabili den bigarren ezaugarri aukeraketa metodoa da. Metodo hau, metodo multibariatu bat da, hau da kontuan hartzen ditu ezaugarrien artean existitu daitezkeen erlazio edo gainezarpenak. Honen helburua, ezaugarri azpimultzo egokiena aukeratzea da, baina kasu honetan, ezaugarriak banan-banan neurtu beharrean multzoka izan dezaketen baliagarritasuna neurtzea da estrategia. Beraz bada, arazoa optimizazio problema bat moduan ikus daiteke, multzo batetik irizpide bat jarraituz, azpimultzo hoberena lortzen saiatzen den optimizazio problema bat moduan.

Estatistiko honek ezaugarri multzo batzuk beste batzuen aurrean hobesteko bi irizpide hartzen ditu kontuan. Batetik, aukeratu den azpimultzoko ezaugarriak klasearekiko korrelatuak egotea hobesten du. Bestetik, ezaugarri hauek beraien artean informazio gutxi konpartitzea, hau da beraien arteko korrelazioa baxua izatea. Bi irizpide hauek bakar batean konbinatzen dira ondoren erakusten den moduan:

$$h = \frac{k\overline{r_{cf}}}{\sqrt{k + (k-1)\overline{r_{ff}}}} \tag{4.9}$$

k multzoko ezaugarri kopurua, $\overline{r_{cf}}$ multzoko ezaugarriek klasearekiko duten korrelazio batazbestekoa eta $\overline{r_{ff}}$ ezaugarriek beraien artean bikoteka duten korrelazio batazbestekoa izanik.

Korrelazio neurritzat neurri ezberdinak erabili daitezke, gure sisteman aurretik azaldu den Information gain korrelazioa erabiliko da.

Azpimultzo egokienaren aukeraketa *NP-hard* problema bat da, hau da, ezin da denbora polinomikoan burutu. Horregatik, *CFS* teknikek ez dute soluzio hoberena aurkitzen, honen hurbilpen bat baizik. Honetarako algoritmo jaleak(*greedy*) erabiltzea ohikoa da. Gure kasuan 5 adarkatzetako best first algoritmo bat erabili da bilaketa algoritmotzat.

4.2 Sailkapenerako meta-algoritmoak

Meta-algoritmoak, sailkapen algoritmo ohikoen gainean eraikitzen diren algoritmoak dira. Algoritmo hauek sailkapen teknika arrunten emaitzak hobetu edo gure arazora egokitzeko erabili ohi dira. Nolabait, gure problemaren izaera hobeto jasotzeko helburuz erabiltzen dira. Gure arazoak bi berezitasun aurkezten ditu.

Batetik, klase aldagaiaren balioek ordenazio bat jarraitzen dute. Hau da $A1 < A2 < B1 < B2 < C1 < C2$. Klase aldagaia ordenatua denean komuna da erregresio teknikak erabiltzea. Baina gure kasuko klasea ordenatua izateaz gain diskretua da, beraz teknika hauek ez dute zentzurik gure kasuan. Gure moduko kasuetarako bereziki sorturiko *Ordinal Classification* izeneko teknika bat erabili da arazo hau konpontzeko.

Bestetik, gure klaseak beste bitxitasun bat aurkezten du. Testuen mailaketa burutzen duten adituen arteko adostasuna ez da batere handia. Aditu batek A2 moduan sailkatuko lukeen testu bat, beste aditu batek B1 moduan sailka dezake. Honek emaitzak neurtzeko orduan malgutasun baten beharra erakusten du. Hau da, ez da arazo handia A2ko testu bat B1 moduan sailkatzea. Aldiz, arazoa handiagoa da A2ko testu bera B2 moduan sailkatzea edo are handiagoa C1 edo C2 moduan sailkatzen bada. Arazo hau konpontzeko *Cost sensitive* teknikak erabili dira, sailkapen oker batzuei besteei baino garrantzia handiagoa emateko asmotan.

4.2.1 Ordinal classification

Ordinal classification(OC) teknikak klase ezaugarri ordenatu eta diskretua duten sailkapen problemetarako erabiltzen dira. Hauen helburua, klase aldagaiaren ordenazioa aprobetxatzea da, horrela emaitza egokiagoak lortzeko asmotan. Gure kasuan [12] artikuluan aurkezturiko teknika aplikatu da. Jarraian teknika honen funtzionamendua aurkezten da, teknikarekin lorturiko emaitzak ikusi nahi izanez gero jo ebaluazio kapitulura.

Teknikaren funtzionamendua azaltzeko artikulu bertako adibidea erabiliko dugu eta gero gure egoerara eratorriko dugu. Artikulua aukeratu den klase aldagaia *tenperatura* da, baina *tenperatura* modu diskretuan harturik. Adibidean *tenperatura* klaseak hiru balio hartzen ditu, hotza (*cold)*, epela (*mild)* eta beroa (*hot*). Sailkapen arazo hau gurea bezala ordenatua eta diskretua da eta beraz $cold < mild < hot$.

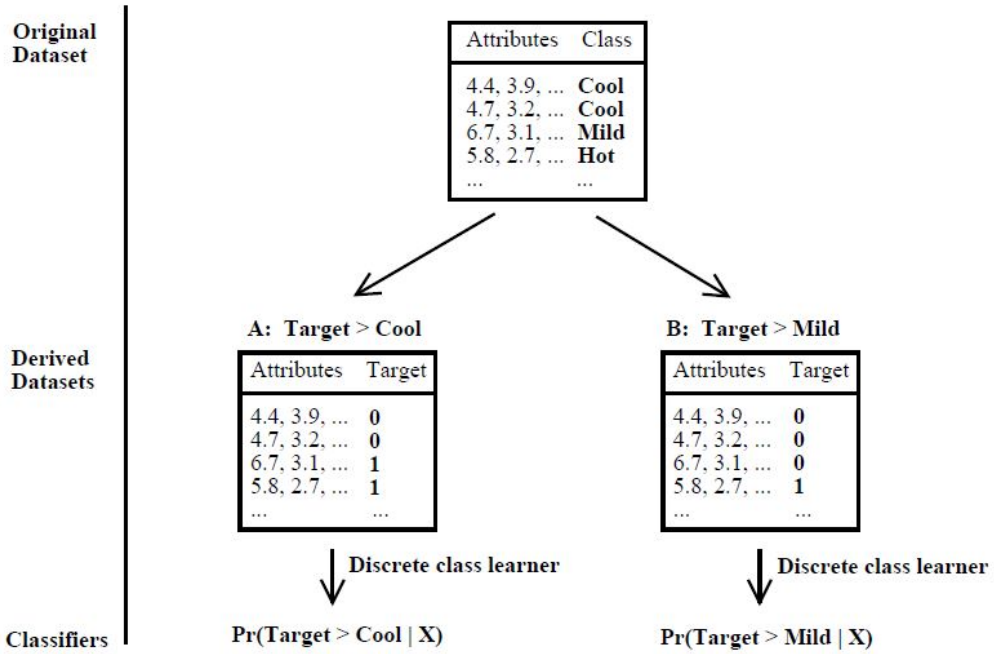
OC teknika hau erabiliz lortu nahi dena klasearen ezaugarriak elkarren artean erlaziorik izango ez balute bezala ez tratatzea da. Horretarako lehenik balio bihurketa bat burutzen da eta N balio posible izatetik $N-1$ balio lortzen dira. Balio berri hauek balio zaharren ondoz-ondoko balioz osaturiko multzo bati erreferentzia egiten diote, horrela ordenatasunaren zentzua mantendu nahian. Adibidera jotzen badugu, bertan bi balio sortuko genituzke, $Target > cold$ eta $Target > mild$. Behin balioak bihurturik balio berri posible bakoitzarentzat sailkatzaile bitar ohiko bat sortuko genuke. Aipaturiko prozesua 4.1 irudian ikus daiteke.

Behin sailkatzaileak entrenaturik, instantzia berrien klasea iragartzea gelditzen zaigu. Honetarako instantzia berria aurretik entrenatu ditugun sailkatzaile guztietan sailkatzen da, sortu berri ditugun balio berrien probabilitateak lortuz. Baina ez dira balio berriak guk nahi ditugunak, beraz datuak berriro transformatu behar ditugu. Demagun sailkatzaileek honako balioak itzuli dizkigutela sailkatzaileek.

- $p(target > cold) = 0.95$
- $p(target > mild) = 0.70$

Hortiko honakoa ondorioztatuko genuke:

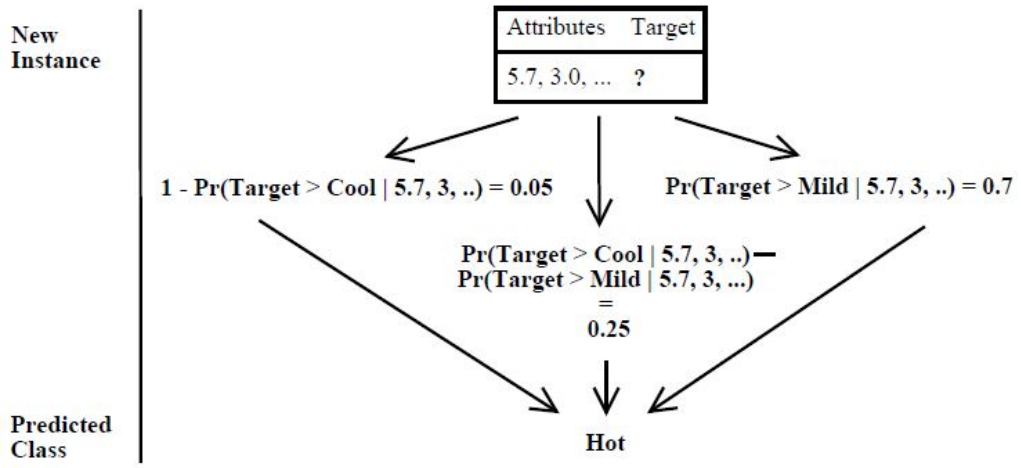
- $p(cold) = 1 - p(target > cold) = 1 - 0.95 = 0.05$
- $p(mild) = p(target > cold) - p(target > mild) = 0.95 - 0.70 = 0.25$



Irudia 4.1: Ikasketa prozesua, [12]

- $p(hot) = p(target > mild) = 0.7$

Behin probabilitate guztiak transformaturik, hemendik handiena aukeratzea soilik beharko genuke, gure iragarpena emateko. Deskribaturiko prozesua 4.2 irudian ikus daiteke.



Irudia 4.2: Iragarpen prozesua, [12]

Gure arazora itzuliz, ikusi berri dugun teknika aplikatzeko honako balio berriak sortu beharko genituzke:

- $target > A1$
- $target > A2$
- $target > B1$
- $target > B2$
- $target > C1$

Behin balio horiek izanik, bakoitzarentzat sailkatzaile bat sortuko genuke eta beren probabilitateak lortzeko erabili. Amaitzeko ondorengo erregelak aplikatuz hasierako balioen probabilitateak lortuko genituzke:

- $p(A1) = 1 - p(target > A1)$
- $p(A2) = p(target > A1) - p(target > A2)$
- $p(B1) = p(target > A2) - p(target > B1)$
- $p(B2) = p(target > B1) - p(target > B2)$
- $p(C1) = p(target > B2) - p(target > C1)$
- $p(C2) = p(target > C1)$

4.2.2 Cost sensitive Learning

Sailkapen problema batzuetan sailkapen errore guztiak ezin dira berdin baloratu. Adibide moduan, medikuntza arlora jotzen badugu, ez ditu kostu berberak pertsona batek minbizia duela iragartzea hau ez duenean, edo pertsona batek minbizirik ez duela iragartzea hau duenean. Batean pertsonak ezusteko bat jasango du baina ez du kalterik jasango, baina bestean pertsona ez da konturatuko minbizirik duenik, baina honen kalteak jasango ditu. Horrelakoak ekiditeko, batzuetan sailkatzailearen asmatze tasa orokor altuago bati uko egin behar izaten zaio, kasu konkretu batzuetako akatsak guztiz minimizatzeko. Hori da hain zuzen cost sensitive teknikek egiten dutena. Asmatze tasa orokorra zertxobait oker dezake batzuetan, baina errore larrienak minimizatzen ditu.

Gure kasura itzuliz klaseko balioen ordenari garrantzia emateko erabili nahi da teknika hau. Hau da, B1 mailako testu baten maila iragarri nahi bada kostu txikia eman nahi zaio A2 edo B2 moduan sailkatua izan bada, baina kostua handituz joango da urrunagoko maila batean sailkatu ahala. Kostu matrize ezberdinekin probak burutu dira, hauek ebaluazio kapituluan ikus daitezke. 4.5 taulan gure gairako erabiltzea posible litzatekeen kostu matrize bat erakusten da adibide moduan. Bertan errorea burutzearen kostua bere benetako klasearekiko distantziarekiko esponentzialki proportzionala da.

	A1	A2	B1	B2	C1	C2
A1	0	1	2	4	8	16
A2	1	0	1	2	4	8
B1	2	1	0	1	2	4
B2	4	2	1	0	1	2
C1	8	4	2	1	0	1
C2	16	8	4	2	1	0

Taula 4.5: Kostu matrize baten adibidea

Bi teknika nagusi burutzen dira deskribaturikoa lortzeko, hauek jarraian azaltzen dira.

Aukeraketa funtzioaren eraldaketa

Sailkatzaile batzuetan nahiko tribiala da kostu funtzio bat gehitzea. Adibide moduan Naive Bayes sailkatzaileak klase balioa iragartzeko orduan probabilitate handieneko balioa aukeratzen du normalean. Probabilitate horiek kostu balioekin konbinatu daitezke, probabilitate handiena duen balioa aukeratu beharrean arrisku posible txikiena duena aukeratzeko. Adibide moduan X balioari 0.7ko probabilitatea eman bazaio eta 2ko kostua badu, honen kostua $(1 - 0.7) * 2$ izango litzateke, hau da, akatsa egiteko probabilitatea bere kostuarekin biderturik. Teknika hau kasu batzuetan soilik da aplikagarria, sailkapen algoritmo batzuk ez baitute probabilitate baliorik erabiltzen.

Entrenamendu datuen eraldaketa

Bigarren teknika hau sailkapen algoritmo guztiekin da aplikagarria. Bertan, ez da sailkapen algoritmoa aldatzen bere horretan, honen entrenamendu datuak baizik. Metodo honek entrenamendu datuen distribuzioa aldatzen du

instantziak errepikatuz, hauen distribuzioa kostu matrizearen distribuzioaren berdin uzteko asmotan. Horrela, kostu handia duten akatsak burutzen dituzten instantziak presentzia handituko dute datuetan, eta beraz ikaste prozesuan sailkapen algoritmoek garrantzia handiagoa emango diete eta hobeto ikasiak izango dira, akats larrien probabilitatea jaitsiz.

5 Kapitulua

Ebaluazioa

Gaien Aurkibidea

Kapitulu honetan, aurretik deskribatu den sistema osoaren ebaluazio bat erakusten da. Analisia bi zatitan banatuko da, batetik, ezaugarrien azterketa bat burutuko da, hauek klasearekiko eman dezaketen informazioa aztertzeko. Bestetik, sistemaren emaitzak bere osotasunean aztertuko dira.

5.1 Esperimenturako datu multzoa

Esperimentua, corpora egoki baten faltaren ondorioz, ez da hizkuntza markoko 6 mailen gainean burutuko. Horren ordez lau maila soilik hartuko dira kontuan, B1, B2, C1 eta C2. Maila bakoitzerako 300 testu erabiliko dira garapenerako eta 41 bukaerako testerako, guztira 1200 + 164 testuko bilduma bat osatuz. Bilduma honen jatorri zein ezaugarriei buruz informazio gehiago nahi izanez gero ikus 2. kapitulua.

5.2 Ezaugarrien analisia

Atal honetan sistemarako sortu diren ezaugarriak aztertuko dira. Lehenik ezaugarriak modu orokorrean aztertuko dira, sistema orokorrean hauek duten garrantzia aztertzeko. Ondoren, ezaugarriak irakurmen maila bakoitzean duten zeresana aztertuko da.

5.2.1 Analisi orokorra

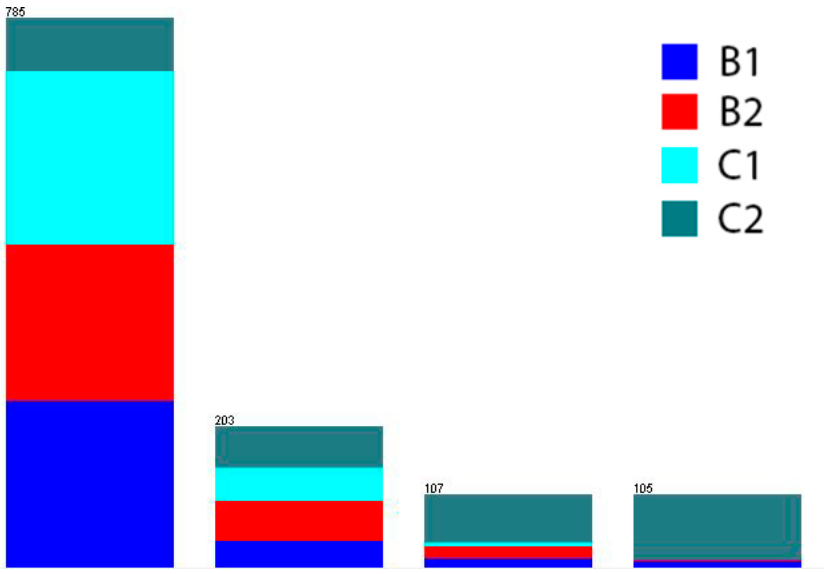
Ezaugarri analisiarekin hasteko ezaugarri guztiak klasearekiko duten esangura neurtu da. Honetarako klasearekiko informazio irabazia (ikus 4.1.1 atala) neurtu da bakoitzarentzat. Ondoren ezaugarri guztiak informazio irabaziaren arabera ordenatuak izan dira, handitik txikira.

Aipaturiko zerrendan 6300etik 969 ezaugarrik dute 0 baino handiago den informazio irabazia. Bestalde, informazio irabazirik handiena 0.3 baino txikiagoa da, hau, kantitate ertain/txikia kontsideratzen delarik. 5.1 taulan aipaturiko zerrendatik lehen 10 ezaugarriak erakusten dira, hau da klasearekiko informazio irabazi gehien aurkezten duten ezaugarriak. Alde handiz informazio gehien aurkezten duen ezaugarria aditz modalen maiztasuna da. 5.1 irudian ezaugarri hau hurbilagotik ikus dezakegu. Bertan ezaugarriaren diskretizazioaren efektua (informazio irabazia kalkulatu aurretik erabili den diskretizazio berbera) erakusten da. Irudian ezaugarriak balio baxuak hartzen dituenean, hau da, aditz modal maiztasuna txikia denean, informazio handirik ezin dela lortu ikus daiteke. Balio txikietan irakurgarritasun maila guztiak antzeko agertzen baitira. Aldiz, aditz maiztasuna handia denean,

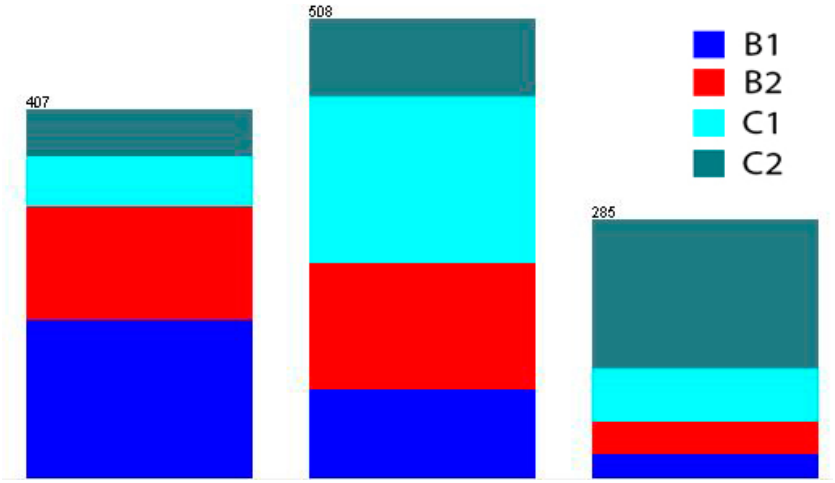
Informazio Irabazia	**Ezaugarria**
0.27807	Aditz modal maiztasuna
0.20147	Lokailu Esplikatzaile + Lokailu egitura maiztasuna
0.20142	Aditz + Aditz + Aditz egitura maiztasuna
0.19855	Sinonimo aberastasuna
0.19542	Zuhaitz sintaktiko sakonera batazb.
0.1927	Aditz sintagma maiztasuna
0.18055	Lokailu + izen egitura maiztasuna
0.17576	Aditz + aditz egitura maiztasuna
0.17535	Hitzeko karaktere kopurua batazb.
0.16693	Emendiozko lokailu maiztasuna

Taula 5.1: Lehen 10 ezaugarriak informazio irabaziaren arabera

irakurgarritasun mailen distribuzioa oso ezberdina da. Bertan C2 maila nabarmentzen da, hau da C2 mailan aditz modalen agerpen maiztasuna handiagoa dela esan dezakegu. Fenomeno hau Aditz + Aditz + Aditz egituraren maiztasuna errepresentatzen duen ezaugarrian ere ikus dezakegu 5.2 irudian ikus daitekeen moduan. Hiru aditz jarraian agertzea egitura konplexua da itxura eta horregatik ematen da gehiago irakurmen maila altua duten testutan.

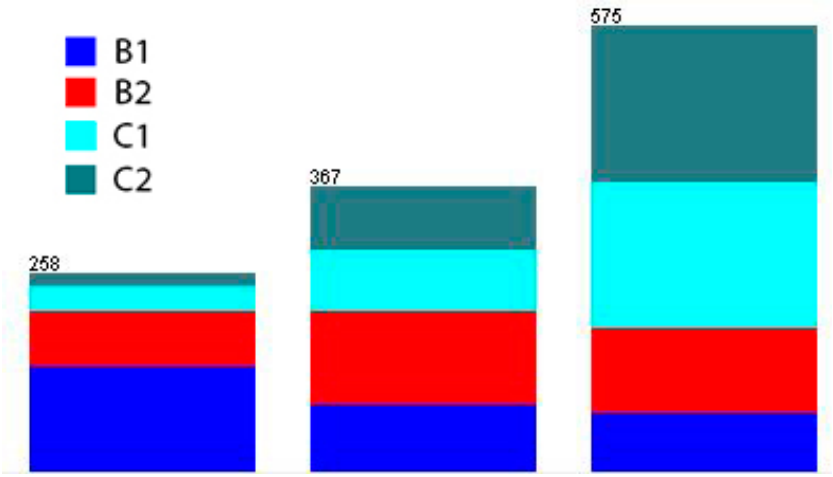


Irudia 5.1: Aditz modal maiztasunaren distribuzioa diskretizatuta.

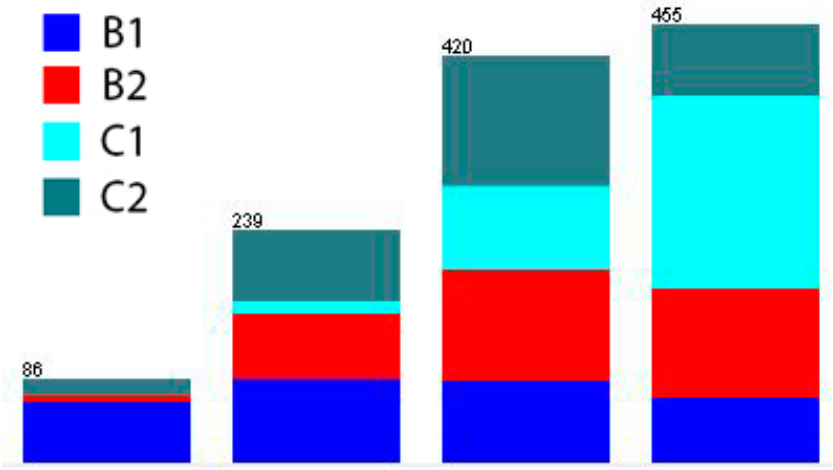


Irudia 5.2: Aditz + Aditz + Aditz egituraren maiztasunaren distribuzioa diskretizatuta.

Sistemarako sortu diren aldagai elaboratuenei erreparatzen badiegu, hauetatik bi lehen 10 ezaugarritan agertzen direla ikus dezakegu, ezaugarri hauen baliozkotasuna erakutsiz. Sinonimoen erabilpen aberatsa errepresentatzeko sorturiko ezaugarria laugarren ezaugarririk esanguratsuena da eta zuhaitz sintaktikoaren sakonera errepresentatzen duen ezaugarria bosgarrena. Ezaugarri hauen diskretizazioak 5.3 eta 5.4 irudietan ikus daitezke. Badirudi, ezaugarri hauekin izandako intuizioa betetzen dela, hau da zuhaitz sakonera txikia duten testuek maila baxuko testu izateko joera dute. Sinonimo aberastasun balioarekin ere antzeko fenomeno bat gertatzen da, sinonimo aberastasuna oso txikia denean probabilitate handiz testua maila baxukoa da.



Irudia 5.3: Zuhaitz sakoneraren batazbestekoaren distribuzioa diskretizatuta.



Irudia 5.4: Sinonimo aberastasunaren distribuzioa diskretizatuta.

Zerrenda bera ikuspegi linguistikoago batetik begiratuz, ezaugarriak bere funtzio linguistikoaren arabera sailkatu eta multzokatuak izan dira. Multzo bakoitzarentzat bere barneko ezaugarrien informazio irabaziaren batazbestekoa kalkulatu da. Informazio hau 5.2 taulan ikus daiteke. Multzo bakoitzaren barneko atributuak zein diren ikusi nahi izanez gero ikus 3 kapitulua.

Ezaugarri multzoa	Bataz besteko informazio irabazia
Lexikoak	0.076
Morfologikoak	0.068
Morfosintaktikoak	0.097
Sintaktikoak	0.083
Pragmatikoak	0.093

Taula 5.2: Ezaugarriak linguistikoki multzokaturik

Taulari erreparatuz, maila morfosintaktiko zein pragmatikoan aurkitzen diren ezaugarriak esangura handiena dutela dirudi testu baten irakurgarritasun maila zehazterakoan. Aldiz, ezaugarri morfologiko zein lexikoek ez dutela horrenbesteko garrantzirik dirudi.

Oharra: Aldagai batek izan dezakeen korrelazio eta beraz informazio irabazi maximoa aldagaiak har ditzakeen balio posible kopuruaren araberakoa da. Horregatik balio posible kopuru ezberdinak dituzten aldagaien informazio irabazien batazbestekoa kalkulatzea ez da guztiz zehatza. Hala ere, aldagai bakoitzak, diskretizazio ondoren, har ditzakeen balio kopuruak begiratuz, hauek ez direla oso ezberdinak ikusi da, eta beraz arazo honek datuetan sor ditzakeen desbideraketak arbuiagarritzat jo dira gure kasurako.

5.2.2 Mailakako analisia

Sistema osoan ezaugarriek duten esangura neurtu ostean, kapitulu honetan ezaugarriek maila bakoitzarekiko duten zeresana aztertuko da. Horretarako aurreko entrenamendu datu berberak erabiliko dira aldaketa bakar batekin. Klase aldagaia bitar bihurtuko da, maila hori den hala ez adierazteko. Horrela maila hori besteengandik ezberdintzeko beharrezko aldagaiak aztertuko dira, maila bakoitzean zein ezaugarri diren garrantzitsuak ikusi ahal izateko.

5.2.2.1 B1 Maila

5.3 taulan B1 maila beste mailetatik bereizteko 10 ezaugarri esanguratsuenak ikus daitezke. Bertan hitz kategoriarekin loturiko pare bat ezaugarri aurki ditzakegu, adjektibo maiztasuna zein aditz trinkoena. Hitzen batazbesteko luzerak maila honetan zeresan handia duela dirudi eta zuhaitz sintaktikoen sakonera ere garrantzizkoa da. Bestalde, dependentzia sintaktiko ugariren egituren maiztasunak ere aurki ditzakegu kopuruan handian. Guztira 938 ezaugarrik dute 0 baino handiagoa den informazio irabazi bat.

Informazio Irabazia	Ezaugarria
0.09942	Adjetibo maiztasuna
0.09594	Hitzeko karaktere kopurua batazb.
0.09128	ncsubj > cmod dependentzia egituraren maiztasuna
0.08653	Zuhaitz sintaktikoen sakonera
0.08475	lot > ROOT dependentzia egituraren maiztasuna
0.08298	auxmod > lot dependentzia egituraren maiztasuna
0.08294	Genitibo kasu markaren maiztasuna
0.07766	ncmod > cmod dependentzia egituraren maiztasuna
0.07765	auxmod > ROOT dependentzia egituraren maiztasuna
0.07716	Aditz trinko maiztasuna

Taula 5.3: Lehen 10 ezaugarriak informazio irabaziaren arabera. (B1)

Datuak ikuspegi linguistikoago batetik helduz 5.4 taulan ezaugarriak beren funtzio linguistikoaren arabera multzokaturik ikus ditzakegu. Bertan ezaugarri multzo bakoitzarentzat bere ezaugarrien informazio irabazien batazbestekoa kalkulatu da. Beste zutabe batean, orokorrean multzo berak zuen informazio irabazia erakusten da, konparaketak egin ahal izateko.

Ezaugarri multzoa	I.I. B1	I.I. (Orokorra)	Aldaketa
Lexikoak	0.039	0.076	-0.037
Morfologikoak	0.037	0.068	-0.031
Morfosintaktikoak	0.044	0.097	-0.053
Sintaktikoak	0.042	0.083	-0.041
Pragmatikoak	0.035	0.093	-0.058

Taula 5.4: Ezaugarriak linguistikoki multzokaturik beren informazio irabaziarekin(I.I.) (B1)

Badirudi, B1 mailan ezaugarri pragmatiko zein morfosintaktikoek garrantzia txikiagoa dutela normalean baino. Hala ere ezaugarri morfosintaktikoak multzo guztietatik garrantzitsuenak izaten jarraitzen duten. Aldiz, ezaugarri pragmatikoak informazio irabazi txikiena duten taldea izatera pasa dira.

5.2.2.2 B2 Maila

B2 mailan esperimentu bera burutuz, datuek atentzioa deitzen dute. 6300tik soilik 198 ezaugarrik dute nulua ez den informazio irabazi bat eta informazio

irabaziok orokorrean txikiak dira. Honek, etorkizunean maila honekin arazoak aurkituko ditugula erakusten digu. Dirudienez, maila hau oso lausoa da eta beraz besteengandik ezberdintzeko arazoak egongo dira sailkapen fasean.

5.5 taulan zerrenda honetako lehen 10 posizioak ikus ditzakegu. Lehen begirada, ezaugarri zerrendak B1ekoarekin zerikusi txikia duela nabaritu dezakegu. Maila honetan dependentzia sintaktikoek zeresan txikia dutela dirudi, B1 maila ez bezala, hemen ez baitago horrelako ezaugarririk top 10ean. Aldiz, hitzen kategoriek eta berauen egituraketek garrantzia handia hartzen dute maila honetan.

Informazio Irabazia	**Ezaugarria**
0.03822	Lokailu esplikatioen maiztasuna
0.03402	Kausazko lokailuen maiztasuna
0.03305	Aditz modal maiztasuna
0.02775	Determinatzaile + lokailu + aditz egituraren maiztasuna
0.02749	Hitz luzeera batazbestean
0.02713	Aditz trinko + aditz + aditz egituraren maiztasuna
0.02703	Adberbio + aditz trinko egituraren maiztasuna
0.02599	Adjektibo + aditz trinko + lokailu egituraren maiztasuna
0.02582	Aditzlagun + adberbio+ adberbio egituraren maiztasuna
0.02430	Izen berezi + lokailu egituraren maiztasuna

Taula 5.5: Lehen 10 ezaugarriak informazio irabaziaren arabera. (B2)

Ezaugarri multzoei dagokionez informazio irabazi txikiaren fenomenoa nabaria da 5.6 taulan. Informazio kantitate guztiak asko jaitsi dira, balio orokorrekin konparaketa egiteak zentzurik ez duen punturaino. Hala ere, datuak modu erlatiboan begiratuz nabaria da aurretik aipaturiko ezaugarri sintaktikoen informazioaren beherakada. Hauen esangura maila honetan kasi nulua da. Aldiz, ezaugarri pragmatikoek, B1 mailan ez bezala, maila honetan garrantzia handia dutela dirudi.

Ezaugarri multzoa	I.I. B2
Lexikoak	0.003
Morfologikoak	0.002
Morfosintaktikoak	0.006
Sintaktikoak	0.000
Pragmatikoak	0.010

Taula 5.6: Ezaugarriak linguistikoki multzokaturik beren informazio irabaziarekin(I.I.) (B2)

5.2.2.3 C1 Maila

C1 mailan informazio irabaziaren balioak ohikoagoak dira. Bertan 6300etik 704 ezaugarrik dute nulua ez den informazio kantitate bat. Maila honetan kategoria ezaugarriek garrantzia izaten jarraitzen dute, baina bestelako ezaugarri batzuk ere agertzen hasten dira. Hala nola, sinonimo aberastasunaren ezaugarria, ezaugarri esanguratsuena da maila honetan. Bestalde, lokailuei erreferentzia egiten dieten ezaugarriak agertzen hasten dira, testuen organizazio eta kohesio altuago baten adierazle.

Informazio Irabazia	Ezaugarria
0.12951	Sinonimo aberastasuna hitzekiko normalizatua
0.0935	Aditz + izen + aditz trinko egituraren maiztasuna
0.09335	Sinonimo aberastasun kontzeptu kopuruarekiko normalizatua
0.07339	Aditz modalen maiztasuna
0.07264	Aditz + aditz + izen egituraren maiztasuna
0.06941	HAOS kategoriaren maiztasuna
0.06134	Emendiozko lokailuen maiztasuna
0.05713	Determinatzaile+ izen + aditz egituraren maiztasuna
0.0567	Adjetibo + determinatzaile + izen egituraren maiztasuna
0.05531	lot > Root dependentzia egituraren maiztasuna

Taula 5.7: Lehen 10 ezaugarriak informazio irabaziaren arabera. (C1)

Ezaugarri multzoei dagokienez, 5.8 taulan ikus dezakegunez, multzo guztiek nahiko modu orekatuan dute esangura maila honetan. Sintaxiak bere garrantzia berreskuratzen du B2 mailarekin alderatuz eta maila honetako multzorik esanguratsuena morfosintasixa den arren, nahiko modu hurbilean

agertzen dira beste multzoak. Multzo pragmatikoa da, balio orokorrarekiko gehien galtzen duen multzoa.

Ezaugarri multzoa	I.I. C1	I.I. (Orokorra)	Aldaketa
Lexikoak	0.022	0.076	-0.044
Morfologikoak	0.023	0.068	-0.045
Morfosintaktikoak	0.027	0.097	-0.070
Sintaktikoak	0.023	0.083	-0.060
Pragmatikoak	0.020	0.093	-0.073

Taula 5.8: Ezaugarriak linguistikoki multzokaturik beren informazio irabaziarekin(I.I. (C1))

5.2.2.4 C2 Maila

C2 mailan 6300 ezaugarrietatik 938k nulua baino informazio irabazi handiagoa erakusten dute. 5.9 taulan ikus dezakegun moduan maila honetan lokailuen esangura handia da. Lehen 10 ezaugarrien artean kausazko, emendiozko, moduzko zein lokailu esplikatiboak aurki ditzakegu, aurreko mailetan baino esaldien arteko antolaketa konplexuagoak erakutsiz. Maila honetan, zenbait kategoria eta hauekin sortutako egiturek ere garrantzia nabaria dute.

Informazio Irabazia	Ezaugarria
0.2371	Aditz modalen maiztasuna
0.1864	Lokailu esplikatiboen maiztasuna
0.18378	Emendiozko lokailuen maiztasuna
0.18163	Hitz luzeraren batazbestekoa
0.14625	Kausazko lokailuen maiztasuna
0.13421	Aditz + aditz+ aditz egituraren maiztasuna
0.13052	Aditz + izen + aditz trinko egituraren maiztasuna
0.10802	Lokailu + izen egituraren maiztasuna
0.10323	Moduzko lokailuen maiztasuna
0.10274	Aposizioa kopurua izen sintagma kopuruarekiko

Taula 5.9: Lehen 10 ezaugarriak informazio irabaziaren arabera. (C2)

5.10 taulan erakusten den ezaugarrien multzokatzeari erreparatuz, maila morfologikoa da garrantzia txikienekoa maila honetan. Aldiz, ezaugarri multzo pragmatikoa beste guztien gainetik gailentzen da informazio irabaziari

dagokionez. Pragmatikak esaldien arteko egituraketa konplexuak aztertzen ditu, maila honetan ohikoak izan ohi diren egiturak, hortik ezaugarrion balio altua.

Ezaugarri multzoa	**I.I. C2**	**I.I. (Orokorra)**	**Aldaketa**
Lexikoak	0.036	0.076	-0.040
Morfologikoak	0.023	0.068	-0.045
Morfosintaktikoak	0.052	0.097	-0.045
Sintaktikoak	0.051	0.083	-0.032
Pragmatikoak	0.068	0.093	-0.025

Taula 5.10: Ezaugarriak linguistikoki multzokaturik beren informazio irabaziarekin

5.2.3 Aukeratutako ezaugarriak

Ezaugarrien inguruan buruturiko azterketak, eskuartean ditugun datuei buruz informazio gehiago lortzeko balio izan digu. Aurrera jarraitzeko ordea, ezaugarri multzo konkretu bat aukeratu beharra dago. Ezaugarri multzo murriztu bat eta ahal den neurrian emaitza ahalik eta egokienak emango dizkiguna. Honetarako CFS aldagai aukeraketa metodoa erabili da. Metodo honek klasearekiko ezaugarri esanguratsuenak lortzen ditu, beti ere ezaugarrien arteko redundantzia ahalik eta murritzen mantenduz. Metodo honi buruz informazio gehiago nahi izanez gero ikus 4.1.2 atala.

CFS algoritmoa eta Best First bilaketa estrategia erabiliz 314979 azpimultzo ebaluatuak izan dira. Guzti hauetatik CFS balio handiena lortu duen multzoak 0.32 balioa lortu du. Jarraian azpimultzo honetako 56 ezaugarriak zerrendatzen dira. Ezaugarri hauek izango dira ebaluazioaren hurrengo azterketa guztietan erabiliko direnak, besterik esan ezean.

- Aditz modal maiztasuna
- Izen elipsi maiztasuna
- Aposizio kopurua izen sintagma kopuruarekiko
- cmod > ncsubj dependentzia egituraren maiztasuna
- gradmod > ncsubj dependentzia egituraren maiztasuna
- ncpred > ncsubj dependentzia egituraren maiztasuna

- prtmod > aponcmod dependentzia egituraren maiztasuna
- Ergatibo kasuaren maiztasuna
- Genitibo kasuaren maiztasuna
- Hitzen luzera batazbestekoa
- Aurkako lokailu maiztasuna
- Emendiozko lokailu maiztasuna
- Lokailu esplikatibo maiztasuna
- Kausazko lokailu maiztasuna
- Moduzko lokailu maiztasuna
- Paritibo kasu maiztasuna
- Aditz + Aditz egituraren maiztasuna
- Aditz + Aditzlagun egituraren maiztasuna
- Adjektibo + Interjekzio egituraren maiztasuna
- Aditzlagun + Lokailu egituraren maiztasuna
- HAOS + Aditz trinko egituraren maiztasunaT
- Izenordain + Aditz Izen maiztasuna
- Izen + Adjektibo egituraren maiztasuna
- Izen + Aditz trinko egituraren maiztasuna
- Izen + Izen berezi egituraren maiztasuna
- Lokailu + Izen egituraren maiztasuna
- Adberbio + Adberbio + Lokailu egituraren maiztasuna
- Adberbio + Lokailu + Partikula egituraren maiztasuna
- Aditz izen + Aditz trinko + Izen egituraren maiztasuna
- Aditz izen + HAOS + Aditz trinko egituraren maiztasuna
- Aditz izen + Juntagailu + Partikula egituraren maiztasuna

- Aditz + Aditz + Aditz egituraren maiztasuna
- Aditz + Aditz + Izen egituraren maiztasuna
- Aditz + Aditzlagun + Izen egituraren maiztasuna
- Aditz + Izenordain + Izen egituraren maiztasuna
- Aditz + Izen + Aditzlagun egituraren maiztasuna
- Aditz + Aditz + Aditz trinko egituraren maiztasuna
- Adjektibo + Adberbio + Izen egituraren maiztasuna
- Adjektibo + Adjektibo + Juntagailu egituraren maiztasuna
- Adjektibo + Bestelako + Adjektibo egituraren maiztasuna
- Adjektibo + Izen + Lokailu egituraren maiztasuna
- Adjektibo + Juntagailu + Adjektibo egituraren maiztasuna
- Aditz trinko + Determinatzaile + Determinatzaile egituraren maiztasuna
- Determinatzaile + Aditz + Adizlagun egituraren maiztasuna
- Determinatzaile + Aditz trinko + Partikula egituraren maiztasuna
- Izen berezi + Adberbio + Aditz trinko egituraren maiztasuna
- Izen berezi + Aditz + Juntagailu egituraren maiztasuna
- Izen + Aditz + Lokailu egituraren maiztasuna
- Izen + Adjektibo + Aditz egituraren maiztasuna
- Izen + Izen + Bestelako egituraren maiztasuna
- Juntagailu + Adjektibo + Izen egituraren maiztasuna
- Partikula maiztasuna
- Zuhaitz sintaktikoen batazbesteko sakonera
- Sinonimo aberastasuna hitz kopuruarekiko normalizatua
- Sinonimo aberastasuna kontzeptu kopuruarekiko normalizatua
- Esaldien batazbesteko luzera

5.3 Sistemaren analisia

Atala honetan, jada aukeratuta ditugula, sistemaren eraginkortasuna aztertzea izango da helburua. Lehenik sailkapen algoritmo arruntak erabiliz sistemak aurkezten dituen emaitzak aztertuko dira. Ondoren, sailkapen algoritmo hauen gainean meta algoritmo ezberdinak aplikatuko dira. Amaitzeko, behin azterketa guztia burututa dagoelarik, emaitza egokienak eman dituen sailkatzailea aukeratu eta test-erako gorderiko datu multzoarekin probatuko da.

5.3.1 Emaitza orokorrak

Eskuartean dugun datu multzoa sailkapen algoritmo ezberdinekin probatu da, emaitza onenak hauetariko zeinek ematen dituen ikusteko. Honetarako, garapenerako datuak hartu eta 10 aldiz ikasketa eta test datutan banatu dira. Banaketa hau ausaz burutu da bakoitzean instantzien $\frac{2}{3}$ ikasketarako gordez eta $\frac{1}{3}$ testerako. Burututako test bakoitzeko asmatze datuak gorde dira hauekin batazbesteko asmatze tasa zein desbideratze estandarra kalkulatzeko. Gainera, datu berdinak aprobetxatuz algoritmo bakoitzarentzat T-Test parekatu bat burutu da Naive Bayes gausiar algoritmoa konparaketa oinarritzat hartuz. T-test parekatua burutzeko 0.05eko p-value edo konfidantza maila bat hartu da. Test parekatu bat erabiltzea erabaki da entrenamendurako datuak sailkapen algoritmo bakoitzarentzat berdinak izan baitira. 5.11 taulan esperimentu honen emaitzak ikus daitezke.

Algoritmoa	**Asmatze-tasa**	**T-test(0.05)**
Naive Bayes gausiarra	57.33 ± 2.02	Oinarria
Naive Bayes multinomiala	61.69 ± 2.48	Estatistikoki berdina
Sare bayesiarra, 3 guraso max.	60.29 ± 2.07	Estatistikoki berdina
J48	51.99 ± 2.52	Estatistikoki okerragoa
1NN	51.81 ± 2.05	Estatistikoki okerragoa
5NN	56.00 ± 1.02	Estatistikoki okerragoa
Erregresio logistikoa	58.90 ± 1.83	Estatistikoki berdina

Taula 5.11: Sailkapen algoritmo ezberdinen asmatze tasak eta konparazio estatistikoa

5.11 taula ikus daitekeenez, naive bayes multinomiala da batazbesteko asmatze tasa altuena duena %61.69ko asmatze tasa batekin. Honengandik hurbil 3 guraso maximoko sare bayesiarra aurkitzen da, %60.29ko asmatze

tasarekin. Honek atentzioa dei dezake, normalean sare bayestar batean guraso posible gehiago izatean emaitzak hobetu edo berdintzen baititu. Honen arrazoia ezaugarriak aukeratzeko metodoan aurki daiteke. CFS algoritmoak ezaugarrien arteko korrelazioak minimizatzen ditu eta honek kalte egiten dio guraso anitzeko sare bayestar bati, bere helburua ezaugarrien arteko korrelazio altuak aurkitzea baita. Emaitza okerrenak KNN zein J48 algoritmoek eman dituzte.

Begi bistara, guraso bakarreko sare bayesiarrak hoberena dirudien arren, estatistikoki ezin dezakegu hala kontsideratu. T-testaren arabera bi naive bayes algoritmoen, sare bayesiarraren zein erregresio logistikoaren emaitzak berdinak dira. Hau da, 0.05eko p-value bat ezarriz, ez dago konfidantza nahikoa lau algoritmo hauetatik inork besteak baino emaitza hobeak eman dituela esateko. Beraz bada, eta soilik proba honetan batazbestean hobea izan delako naive bayes multinomiala aukeratu da hurrengo azterketarako.

Klasea	**TP ratioa**	**FP ratioa**	**Doitasuna**	**Estaldura**	**F-Measure**
B1	0,533	0,140	0,559	0,533	0,546
B2	0,430	0,188	0,433	0,430	0,431
C1	0,750	0,184	0,575	0,750	0,651
C2	0,723	0,009	0,964	0,723	0,827

Taula 5.12: Naive Bayes multinomiala, estatistikoak mailakaturik

5.12 taulan aukeratu dugun guraso naive bayes multinomialak emandako emaitzen informazio gehiago ikus daitezke. Taula mailaka ikusiz C2 mailak atentzioa deitzen du bere emaitza onengatik. 0.964ko prezisioa eta 0.09ko false positive ratioa beste mailengandik nabarmentzen dira. Badirudi, maila hau besteengandik bereziki ezberdina dela eta horregatik emaitza onak lortzen dira bertako testuetan. Aldiz, beste muturrean B2 maila aurki dezakegu. Badirudi, ezaugarri aukeraketa fasean buruturiko hipotesia bete dela. Bertan ezaugarriek maila honekiko zuten informazio irabazia oso txikia zela ikusi genuen, eta dirudienez honek maila honetako emaitzetan zeresan handia izan du.

	Iragarri B1	**Iragarri B2**	**Iragarri C1**	**Iragarri C2**
B1	160	96	43	1
B2	80	129	90	1
C1	8	61	225	6
C2	38	12	33	217

Taula 5.13: Naive Bayes multinomiala, Konfusio matrizea

5.13 taulan burutu den sailkapenaren konfusio matrizea erakusten da. Bertan, instantzia bakoitza nola sailkatua izan den ikus daiteke modu laburtu batean. Zutabetan instantziak nola iragarri diren ikus daiteke, eta lerroetan instantzien klase erreala zein den. Diagonal nagusian aurkitzen diren zenbakiek zuzen iragarri diren instantziak errepresentatzen dituzte eta beste zenbakiek erroreak. Hemen ere aurretik aipaturiko bi fenomenoak ikus ditzakegu, C2 klasearen asmatzeak besteenak baino nabari hobeak dira eta B2 klaseak sortzen ditu arazo gehien. Taulan ikus dezakegun moduan B2 klasea nolabait tartean duten iragarpen oker guztiak nahiko handiak dira.

Arlo honetako adituen arteko desadostasuna kontuan izanik arloko literaturan nahiko ohikoa da *adjacent accuracy* deituriko neurria erabiltzea asmatze tasak errepresentatzeko. Neurri hau ordenaturik aurkitzen diren klaseekin da soilik erabilgarria eta bere helburua problemaren lausotasuna kontuan hartzea da. Bere funtzionamendua sinplea da, asmatze tasak berdin kalkulatzen dira, baina 1eko distantziara aurkitzen diren erroreak zuzentzat jotzen dira. Adibide moduan, B2 klasea den testu bat B1 edo C1 moduan sailkatua izan bada asmatze tasa kalkulatzerakoan zuzentzat jotzen da.

5.14 taulan aipaturiko neurriaren datuak erakusten dira, [11] artikuluan frantsezerako sorturiko antzeko sistema baten emaitzekin batera. Argi utzi beharra dago, emaitzak ez direla guztiz konparagarriak artikuluan A1 eta A2 maila ere kontuan hartzen baita sailkapenerako.

	B1	B2	C1	C2	Batazb.
Sortutako Sistema	%85.3	%99.96	%97.33	%83.33	%91.48
T. Francois et al.	%67	%71	%86	%83	%77

Taula 5.14: Naive Bayes multinomiala, Adjacent accuracy

5.3.2 Meta algoritmoen analisia

Atal honetan bi estrategia ezberdin aplikatuko dira sistemaren emaitzak problematikaren naturara egokitzen saiatzeko. Lehenik *ordinal classification* deituriko teknika aplikatuko dugu, honek emaitzetan izan ditzakeen ondorioak aztertzeko. Ondoren *Cost sensitive classification* tekniketaz baliatuko gara bigarren azterketa bat burutzeko.

5.3.2.1 Ordinal classification

Teknika honen helburua klasearen ordinaltasuna aprobetxatuz emaitza egokiagoak lortzen saiatzea da. Hau da, helburu nagusia ez da asmatze tasa inkrementatzea, egiten diren erroreek bere benetako balioarekiko desplazamendu txikiagoa izatea baizik. Teknika honen inguruan gehiago jakiteko ikus 4.2.1 atala.

Teknika honen ondorioak aztertzeko naive bayes multinomialarekin jarraitu da, meta algoritmoa gainetik aplikatuz, horrela emaitzak konparatzeko oinarri bat izatekotan.

Naive Bayes	Naive Bayes + ordinal classification
60.87 ± 3.83	60.52 ± 4.17

Taula 5.15: Ordinal classification vs. naive bayes soilik

5.15 taulan bi sistemen asmatze tasak erakusten dira. Bi sistemen asmatze tasak antzekoa direla ikus dezakegu eta t-testak estatistikoki berdinak direla baieztatzen digu 0.05eko p-value batekin. Beraz asmatze tasei dagokionez ez da alde nabaririk sumatzen.

	Iragarri B1	**Iragarri B2**	**Iragarri C1**	**Iragarri C2**
B1	180	80	39	1
B2	113	98	84	5
C1	18	68	202	12
C2	30	10	33	227

Taula 5.16: Konfusio matrizea (Ordinal classification)

5.16 taulako konfusio matrizeari erreparatzen badiogu, emaitzetan aldaketa batzuk gertatu direla ikusiko dugu. Muturretako asmatze tasek igoera

kontsideragarria jasan dute. Aldaketa honek ordea B2 eta C1 mailetako emaitzak okertzea ekarri du. Aldaketa horiez aparte, ez da aparteko onurarik nabaritzen datuetan, erroreak ez dira hurbilago kokatzen ezta gutxiagotan gertatzen ere. [12] artikuluan aipatzen den moduan, *ordinal classification* teknikek hobekuntzak sortzen dituzte kasu askotan. Aldiz, inolako efekturik sortzen ez duten kasuak ere existitzen dira, eta dirudienez, gure kasua da horietako bat.

5.3.2.2 Cost sensitive classification

Bigarren teknika honen helburua aurrekoaren antzekoa da. Hau da, helburu nagusia ez da asmatze tasa hobetzea, erroreen larritasuna murriztea baizik. Teknika honen oinarria errore mota batzuen kostua markatzean datza. Kostu hauek modeloa eraikitzerakoan kontuan izango dira. Teknikari buruz informazio gehiago nahi izanez gero ikus 4.2.2 atala.

	Iragarri B1	**Iragarri B2**	**Iragarri C1**	**Iragarri C2**
B1	0	1	10	1000
B2	1	0	1	10
C1	10	1	0	1
C2	1000	10	1	0

Taula 5.17: Kostu matrizea

Esperimentua burutzeko erabili den kostu matrizea 5.17 taulan erakusten da. Kostuak zehazterako orduan, helburu nagusia akats larrienak, hau da, benetako klasetik urruntasuna handiena zutenak, asko penalizatzea izan da. Izan ere akats hauek izango dira errealitatean arazo gehien emango duten akatsak.

Naive Bayes	Naive Bayes + cost sensitive
60.87 ± 3.83	58.79 ± 4.52

Taula 5.18: Cost sensitive vs. Naive Bayes soilik

5.18 taulan experimentuan sailkatzaileak izan duen asmatze tasa erakutsi eta kostu matrizea erabili gabeko sailkatzailearen asmatze tasarekin konparatzen da. Cost sensitive teknika erabiliz begibistara emaitza okerragoa den arren aldaketa ez da oso handia eta t-testaren arabera estatistikoki berdinak

dira 0.05eko p-value batekin. Beraz asmatze tasari dagokionez ez da aldaketarik gertatu cost sensitive teknikak erabilita.

	Iragarri B1	**Iragarri B2**	**Iragarri C1**	**Iragarri C2**
B1	152	106	42	0
B2	67	141	92	0
C1	7	58	233	2
C2	13	34	78	175

Taula 5.19: Konfusio matrizea (Cost Sensitive)

Konfusio matrizeari dagokionez aldaketak nabariak dira 5.13 taularekiko, 5.19 taulan ikusi dezakegun moduan. Akats larrienei emandako penalizazioak bere efektuak izan ditu eta akats hauek kopurua nabari jaitsi da. B1 izan eta C2 moduan sailkaturiko instantziak 0 izatera jaitsi dira 1etik eta C2 izan eta B1 moduan sailkaturikoak 38tik 13ra jaitsi dira. Orokorrean, B1, B2 eta C1 eko asmatze tasak mantendu edo igo egin dira. C2ri dagokionez guztizko asmatze tasan beherakada bat gertatu den arren, hau ez da beherakada garrantzitsua izan sistemarako, sortu diren errore berri gehienak maila bakarreko desplazamendua soilik jasan baitute, eta errore hori guztiz onargarria da gure problematikarako. Beste zenbakiek aldaketa batzuk jasan dituzten arren ez dira modu orokorrean eraginik duten aldaketak, beraz cost sensitive teknikak sisteman duen eragina positiboa dela esan daiteke.

5.3.3 Test-erako datuak

Sistemako atal guztiak bere horretan aztertu ondoren ondorio ezberdinak lortu ditugu. Momentura arte lorturiko konbinazio hoberena **CFS bidez lorturiko aldagaiak + naive bayes multinomiala + cost sensitive meta algoritmoa** dituen konbinazioa izan da, asmatze tasa handiena lortzeaz gain, problemarako emaitza egokituak ere eskaintzen baititu. Beraz bada, momentura arteko sistemarik hoberena aukeraturik, honen bukaerako ebaluazioa burutuko dugu atal honetan. Horretarako, sistemaren sorkuntzan inoiz ikusi ez ukitu diren datuak erabiliko dira. Datu hauek klase bakoitzerako 41 instantziaz osaturik daude, guztira 164 instantziako datu multzo bat osatuz.

5.20 taulan esperimentuaren zenbait estatistika ikus daitezke. Bertan ikus daitekeenez emaitzak garapenerako datuekin lorturikoak baino zerbait oke-

rragoak dira, baina orokorrean berdintsu mantentzen dira. C2ren asmatze tasak nabarmen altua izaten jarraitzen du eta B2 klaseak arazoak ematen jarraitzen du prezisioari begiratzen badiogu. Asmatze tasaren jaitsiera orokorra C1en jaitsieraren ondorioa izan daiteke, klase honen prezisioa nabarmen jaitsi da entrenamendu datuekiko.

Klasea	TP ratioa	FP ratioa	Doitasuna	Estaldura	F-Measure
B1	0,463	0,104	0,633	0,463	0,535
B2	0,390	0,226	0,400	0,390	0,395
C1	0,659	0,330	0,435	0,659	0,524
C2	0,583	0,008	0,933	0,583	0,718
Batazbestekoa	0,517	0,186	0,562	0,517	0,523

Taula 5.20: Estatistikak, testerako datuak

Konfusio matrizeari dagokionez, cost sensitive metodoak bere efektua izan du datu hauetan ere. Akats larrienak kopuru murritzean gertatzen direla ikus dezakegu 5.21 taulan. B1 izanda C2 moduan sailkaturiko instantziak 0 izan dira eta C2 izan eta B1 moduan sailkatuak soilik 1.

	Iragarri B1	Iragarri B2	Iragarri C1	Iragarri C2
B1	19	11	11	0
B2	8	16	17	0
C1	2	11	27	1
C2	1	2	7	14

Taula 5.21: Konfusio matrizea (Test datuak)

Beraz bada, asmatze tasan izandako jaitsiera batzuk albo batera utzirik, sistemak espero bezala jokatu du testeko datuekin buruturiko esperimentuan.

6 Kapitulua

Ondorioak eta etorkizuneko lana

Gaien Aurkibidea

Behin proiektua bukatutzat emanda, honen inguruan gertaturiko gorabehera guztiak analizatzeko momentua da. Jarraian proiektuak izan dituen ondorioak zein etorkizuneko lanak deskribatzen dira.

6.1 Proiektuaren ondorioak

Jarraian proiektuaren garapenetik lortu diren ondorio nagusiak zerrendatzen dira:

- Testu bat irakurri ahal izateko beharrezko hizkuntza maila neurtzeko gai den sistema bat sortu da.
- Sistema A1 eta A2 mailetan probatu ezin izan den arren beste lau mailetan buruturiko esperimentuek literaturako beste sistemen parean aurkitzen dela erakusten dute.
- Sistemarako sortu diren aldagai berriek orokorrean klasearekiko esangura handia erakutsi dute:
  - Dependentzia zein kategoria ngramek sorturiko ezaugarritatik zenbait ezaugarri interesgarri sortu dira, irakurgarritasun maila detektatzeko egitura espezifiko batzuk detektatzeko gai direnak.
  - Zuhaitz sintaktikoaren sakonera zein testuaren sinonimo aberastasuna neurtzeko sorturiko ezaugarriek korrelazio altuak erakutsi dituzte klasearekiko, hauek sortzeko hipotesiak bete direlarik.
  - Testuaren jarraitasun semantikoa neurtzeko sorturiko ezaugarria korrelazio altueneko ezaugarrien artean agertzen ez den arren, bere korrelazioa orokorrean ez da guztiz txarra eta baliozko ezaugarritzat kontsidera daiteke.
- Meta-algoritmoei dagokienez, emaitza ezberdinak lortu dira.
  - Ordinal Classification algoritmoak gure sisteman ez du efektu nabaririk izan.
  - Cost sensitive algoritmoak, aldiz, asmatze tasa orokorrak berdintsu mantendu dituen arren errore motetan aldaketa nabariak sortu ditu. Errore larrienak modu nabarian jaitsi dira, beste datuak antzeko mantenduz.

6.2 Etorkizuneko lana

Jarraian proiektuaren etorkizuneko lanak zerrendatzen dira:

- A1 zein A2 mailetako corpora eskuratzea interesgarria izango litzateke sistemaren ebaluazio osoago bat burutzeko eta literaturako beste sistemekin guztiz konparatzeko.
- Testuen jarraitasun semantikoa neurtzeko beste mota besteko estatistikoak sortzea interesgarria izango litzateke, hitzetan soilik oinarritu beharrean kontzeptu mapetan oinarrituz.
- 300 testu baino gehiago dituen corpora bat eskuratzea ere interesgarria izan liteke, sistemak entrenamendu sakonago batekin emaitza hobeak aurkezten dituen ikusteko.
- Sistemak beste mota batzuetako testuekin nola jokatzen duen ikustea interesgarria litzateke, elkarrizketa motako testuekin adibidez.
- Sistema beste hizkuntza ezberdinetarako moldatzea interesgarria litzateke, hizkuntza batetik bestera testuen irakurgarritasunean efektua duten ezaugarriak aldatzen diren edo ez ikusteko. Honetarako bi hizkuntza ezberdinetan aurkitzen den corpus mailakatu bat lortu beharko litzateke.

I Eranskina

ErreXaileko ezaugarri zerrenda

Gaien Aurkibidea

Jarraian ErreXail sistemako ezaugarri guztiak zerrendatzen dira, bere kodeketarekin.

I..1 Ezaugarri orokorrak

1. RAT_WORDCOUNT_SENTENCECOUNT
   Esaldiko hitz kopurua batazbestean.

2. RAT_PHRASECOUNT_SENTENCECOUNT
   Esaldiko sintagma kopurua batazbestean.

3. RAT_KARAKTEREAK_WORDCOUNT
   Hitzeko karaktere luzera batazbestean.

I..2 Ezaugarri lexikoak

Kategoriak

1. RAT_IZE_WORDCOUNT
   Izen kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_IZB_WORDCOUNT
   Izen berezi kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_ADI_WORDCOUNT
   Aditz kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_ADT_WORDCOUNT
   Aditz trinko kopurua, testuko hitz kopuruarekiko normalizatua.

5. RAT_ADL_WORDCOUNT
   Aditz laguntzaile kopurua, testuko hitz kopuruarekiko normalizatua.

6. RAT_ADIZE_WORDCOUNT
   Aditz izen kopurua, testuko hitz kopuruarekiko normalizatua.

7. RAT_ADJ_WORDCOUNT
   Adjektibo kopurua, testuko hitz kopuruarekiko normalizatua.

8. RAT_ADB_WORDCOUNT
   Adberbio kopurua, testuko hitz kopuruarekiko normalizatua.

9. RAT_LOK_WORDCOUNT
   Lokailu kopurua, testuko hitz kopuruarekiko normalizatua.

10. RAT_JNT_WORDCOUNT
    Juntagailu kopurua, testuko hitz kopuruarekiko normalizatua.

11. RAT_DET_WORDCOUNT
    Determinatzaile kopurua, testuko hitz kopuruarekiko normalizatua.

12. RAT_PRT_WORDCOUNT
    Partikula kopurua, testuko hitz kopuruarekiko normalizatua.

13. RAT_IOR_WORDCOUNT
    Izenordain kopurua, testuko hitz kopuruarekiko normalizatua.

14. RAT_GRAD_WORDCOUNT
    Graduatzaile kopurua, testuko hitz kopuruarekiko normalizatua.

15. RAT_ITJ_WORDCOUNT
    Interjekzio kopurua, testuko hitz kopuruarekiko normalizatua.

16. RAT_BST_WORDCOUNT
    Bestelako kopurua, testuko hitz kopuruarekiko normalizatua.

Pertsona ezaugarriak

1. RAT_NOR_WORDCOUNT
   Nor aditz kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_NOR_NORI_WORDCOUNT
   Nor nori aditz kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_NOR_NORK_WORDCOUNT
   Nor nork aditzkopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_NOR_NORI_NORK_WORDCOUNT
   Nor nori nork aditz kopurua, testuko hitz kopuruarekiko normalizatua.

Laburtzapen ezaugarriak

1. RAT_LAB_WORDCOUNT
   Laburtzapen kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_SIG_WORDCOUNT
   Sinbolo kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_SNB_WORDCOUNT
   Akronimo kopurua, testuko hitz kopuruarekiko normalizatua.

Beste ezaugarri lexikoak

1. RAT_ENTI_WORDCOUNT
   Entitate izendatu kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_ADITZMODALAK_WORDCOUNT
   Entitate izendatu kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_ADITZSEMIMODALAK_WORDCOUNT
   Entitate izendatu kopurua, testuko hitz kopuruarekiko normalizatua.

I..3 Ezaugarri morfologikoak

Kasua

1. RAT_ABL_WORDCOUNT
   Ablatibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_ABU_WORDCOUNT
   Adlatibo bukatuzko kasu kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_ABZ_WORDCOUNT
   Adlatibo bide zuzeneko kasu kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_ALA_WORDCOUNT
   Alatibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

5. RAT_SOZ_WORDCOUNT
   Soziatibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

6. RAT_DAT_WORDCOUNT
   Datibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

7. RAT_DES_WORDCOUNT
   Destinatibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

8. RAT_ERG_WORDCOUNT
   Ergatibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

9. RAT_GEL_WORDCOUNT
   kasu kopurua, testuko hitz kopuruarekiko normalizatua.

10. RAT_GEN_WORDCOUNT
    Genitibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

11. RAT_INE_WORDCOUNT
    Inesibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

12. RAT_INS_WORDCOUNT
    Instrumental kasu kopurua, testuko hitz kopuruarekiko normalizatua.

13. RAT_MOT_WORDCOUNT
    Motibatibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

14. RAT_ABS_WORDCOUNT
    Absolutibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

15. RAT_PAR_WORDCOUNT
    Partitibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

16. RAT_PRO_WORDCOUNT
    Prolatibo kasu kopurua, testuko hitz kopuruarekiko normalizatua.

17. RAT_BNK_WORDCOUNT
    Banatzaile kasu kopurua, testuko hitz kopuruarekiko normalizatua.

18. RAT_DESK_WORDCOUNT
    Deskribatzaile kasu kopurua, testuko hitz kopuruarekiko normalizatua.

I..3.1 Aditz aspektu markak

1. RAT_GERO_WORDCOUNT
   Geroaldi marka kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_BURU_WORDCOUNT
   Burutu marka kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_EZBU_WORDCOUNT
   Ez burutu marka kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_PNT_WORDCOUNT
   Puntukari marka kopurua, testuko hitz kopuruarekiko normalizatua.

I..3.2 Aditz aldiak

1. RAT_LEHENALDIA_WORDCOUNT
   Lehenaldi marka kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_ORAINALDIA_WORDCOUNT
   Orainaldi marka kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_ALEGIAZKOA_WORDCOUNT
   Alegiazko marka kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_GEROALDIARKAIKOA_WORDCOUNT
   Geroaldi arkaiko marka kopurua, testuko hitz kopuruarekiko normalizatua.

I..3.3 Aditz moduak

1. RAT_INDIKATIBOA_WORDCOUNT
   Indikatiboan dauden aditz kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_SUBJUNTIBOA_WORDCOUNT
   Subjuntiboak dauden aditz kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_AHALERA_WORDCOUNT
   Ahaleran dauden aditz kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_INPERATIBOA_WORDCOUNT
   Inperatiboan dauden aditz kopurua, testuko hitz kopuruarekiko normalizatua.

I..3.4 Elipsiak

1. RAT_IZE_IZEELI_WORDCOUNT
   Elipsian dauden izen kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_DET_IZEELI_WORDCOUNT
   Elipsian dauden determinatzaile kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_ADL_IZEELI_WORDCOUNT
   Elipsian dauden aditz laguntzaile kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_ADT_IZEELI_WORDCOUNT
   Elipsian dauden aditz trinko kopurua, testuko hitz kopuruarekiko normalizatua.

5. RAT_ADJ_IZEELI_WORDCOUNT
   Elipsian dauden adjektibo kopurua, testuko hitz kopuruarekiko normalizatua.

6. RAT_ADB_IZEELI_WORDCOUNT
   Elipsian dauden adberbio kopurua, testuko hitz kopuruarekiko normalizatua.

I..4 Ezaugarri morfosintaktikoak

1. RAT_IS_SENTENCECOUNT
   Izen sintagma kopurua, esaldi kopuruarekiko normalizatua.

2. RAT_IS_PHRASECOUNT
   Izen sintagma kopurua, sintagma kopuruarekiko normalizatua.

3. RAT_AS_SENTENCECOUNT
   Aditz sintagma kopurua, esaldi kopuruarekiko normalizatua.

4. RAT_AS_PHRASECOUNT
   Aditz sintagma kopurua, sintagma kopuruarekiko normalizatua.

5. RAT_APOSCOUNT_PHRASECOUNT
   Aposizio kopurua, sintagma kopuruarekiko normalizatua.

I..5 Ezaugarri sintaktikoak

Menderakuntza sintagmak

1. RAT_KONPL_WORDCOUNT
   Konpletibozko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_ERLT_WORDCOUNT
   Erlatibozko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_DENB_WORDCOUNT
   Denborazko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_MOD_WORDCOUNT
   Moduzko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

5. RAT_KAUS_WORDCOUNT
   Kausazko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

6. RAT_KONT_WORDCOUNT
   Kontsekutibozko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

7. RAT_BALD_WORDCOUNT
   Baldintzazko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

8. RAT_HELB_WORDCOUNT
   Helburuzko sintagma kopurua, testuko hitz kopuruarekiko normalizatua.

I..6 Ezaugarri pragmatikoak

Juntagailuak

1. RAT_JNT_EMEN_WORDCOUNT
   Emendiozko juntagailu kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_JNT_HAUT_WORDCOUNT
   Juntagailu hautakari kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_JNT_AURK_WORDCOUNT
   Juntagailu aurkari kopurua, testuko hitz kopuruarekiko normalizatua.

Lokailuak

1. RAT_LOK_EMEN_WORDCOUNT
   Emendiozko lokailu kopurua, testuko hitz kopuruarekiko normalizatua.

2. RAT_LOK_AURK_WORDCOUNT
   Aurkako lokailu kopurua, testuko hitz kopuruarekiko normalizatua.

3. RAT_LOK_ESPL_WORDCOUNT
   Lokailu esplikatibo kopurua, testuko hitz kopuruarekiko normalizatua.

4. RAT_LOK_KAUS_WORDCOUNT
   Kausazko lokailu kopurua, testuko hitz kopuruarekiko normalizatua.

5. RAT_LOK_ONDO_WORDCOUNT
   Ondoriozko lokailu kopurua, testuko hitz kopuruarekiko normalizatua.

6. RAT_LOK_MOD_WORDCOUNT
   Moduzko lokailu kopurua, testuko hitz kopuruarekiko normalizatua.

7. RAT_LOK_KONT_WORDCOUNT
   Lokailu kontsezibo kopurua, testuko hitz kopuruarekiko normalizatua.

8. RAT_LOK_BALD_WORDCOUNT
   Baldintzazko lokailu kopurua, testuko hitz kopuruarekiko normalizatua.

9. RAT_LOK_HAUT_WORDCOUNT
   Lokailu hautakari kopurua, testuko hitz kopuruarekiko normalizatua.

Bibliografia

[1] Euskarazko wikipedia. URL: `http://eu.wikipedia.org/`.

[2] Ikasbil. URL: `http://www.ikasbil.net/`.

[3] Ixa ikerketa taldea. URL: `http://ixa.si.ehu.es`.

[4] Jsoup html parser. URL: `http://jsoup.org/`.

[5] E. Agirre (1), X. Artola (1), A. Diaz de Ilarraza (1), G. Rigau(1), A. Soroa (1), and W. Bosma (2). Kaf: Kyoto annotation framework. *IXA group, University of the Basque Country (1), Computational Lexicology and Terminology Lab, VU Amsterdam (2).*

[6] I. Aduriz, M. Aranzabe, J.M. Arriola, A. Diaz de Ilarraza, K. Gojenola, M. Oronoz, and L. Uria. A cascaded syntactic analyser for basque. computational linguistics and intelligent text processing. *pp 124-135.*, 2004.

[7] Aitzol Astigarraga, Koldo Gojenola, Kepa Sarasola, and Aitor Soroa. *TAPE Testu-analisirako PERL erremintak.* Udako Euskal Unibertsitatea (UEU), Bilbo, Spain, 2009.

[8] Ion Madrazo Azpiazu. Hizkuntzaren prozesamendurako teknikak irakaskuntza arloan: galdera sortzaile automatikoa. *Ixa ikerketa taldea*, 2013.

[9] Furnas G. Landauer T. Deerwester S., Dumais S. and Harshman R. Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*, pages 391—407, 1990.

[10] Usama Fayyad and Keki Irani. Multi-interval discretization of continuous-valued attributes for classification learning. 1993.

[11] Thomas François and Cédrick Fairon. An ai readability formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. Association for Computational Linguistics, 2012.

[12] Eibe Frank and Mark Hall. *A simple approach to ordinal classification*. Springer, 2001.

[13] Itziar Gonzalez-Dios, Marıa Jesús Aranzabe, Arantza Dıaz de Ilarraza, and Haritz Salaberri. Simple or complex assessing the readability of basque texts. In *Proceedings of COLING*, volume 2014, 2014.

[14] Mark A Hall and Lloyd A Smith. Practical feature subset selection for machine learning. 1998.

[15] Elisabete Pociello Irigoyen. Euskararen ezagutza-base lexikala: Euskal wordnet. *Euskal Filologian Doktore titulua eskuratzeko aurkezturiko Tesia*, 2007.

[16] Bengoetxea K. and Gojenola K. Application of diferent techniques to dependency parsing of basque first workshop on statistical parsing of morphologically rich languages. *SPMRL and NAACL Workshop Los Angeles*, 2010.

[17] Dennis S. Landauer T.K., McNamara D.S. and Kintsch W. Handbook of latent semantic analysis. In *Mahwah NJ*. Lawrence Erlbaum Associates, 2007.

[18] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.