VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Euskarazko lehen C1 ebaluatzaile automatikoa

Ekhi Azurmendi, Xabier Arregi
eta *Oier Lopez de Lacalle*

125-132 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.15

Euskarazko lehen C1 ebaluatzaile automatikoa

Ekhi Azurmendi[1], Xabier Arregi[1], Oier Lopez de Lacalle[1]

HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU

ekhi.azurmendi@ehu.eus

Laburpena

Artikulu honetan euskarazko idazlanek C1 maila duten edo ez zehazten duen ebaluatzaile automatiko bat garatu dugu. Sistema elikatzeko HABE eta HiTZ arteko hitzarmenaren bitartez lortutako transkribatutako 10.000 idazlan erabili dira. Idazlanen gaiak eduki dezaketen eragina aztertzeko entrenamenduak bi eratara diseinatu ditugu, epealdi bakarreko testuak bakarrik erabilita eta bi epealdietakoekin. Oinarri lerroak finkatzeko euskarazko bi Hizkuntza Eredu (HE), RoBERTa eta Latxa, ereduak entrenatu ditugu, eta ondoren datu eskasiari aurre egiteko, sistemaren gaindoitzea ekiditeko eta errendimendua hobetzeko teknika ezberdinak landu: EDA, SCL eta erregulazioa. Azkenik, sistema ezberdinen portaeren analisiak burutu ditugu, ereduen kalibrazioa eta artefaktuen eragina neurtzeko.

Hitz gakoak: adimen artifiziala, ikasketa automatikoa, hizkuntza ereduak, idazlanen kalifikazio automatikoa

Abstract

In this article, we have developed an automatic evaluator that determines whether texts written in Basque meet the C1 level. To train the system, we used 10,000 transcribed essays obtained through an agreement between HABE and HiTZ. To analyze the potential impact of essay topics, we designed the training in two ways: using texts from only one exam period and using texts from two exam periods. To establish baselines, we trained two Language Models for Basque, RoBERTa and Latxa, and then worked on different techniques to address data scarcity, prevent system overfitting, and improve performance: EDA, SCL, and regularization. Finally, we conducted analyses of different system behaviors to measure model calibration and the impact of artifacts.

Keywords: *artificial intelligence, machine learning, language models, automatic essay evaluation*

1 Sarrera eta motibazioa

Gaur egun hizkuntza teknologiak geroz eta erabiliagoak dira, ChatGPT, Alexa edota Siri bezalako adimen artifizialak oso zabalduta daude gizartean. Teknologia hauek baliabide ugariko hizkuntzetan errendimendu altua izan ohi dute, baina arazoak izaten dituzte euskara bezalako hizkuntza txikiagoekin. Euskarak mundu digitalean presentzia edukitzea garrantzitsua da hizkuntza teknologiak euskarara ekartzeko.

Euskaraz teknologia hauek garatzeko eta beharrezko datuak biltzeko IKER-GAITU proiektua sortu zen 2023an. Artikulu hau IKER-GAITUren barnean kokatzen da, zehazki idatzizko euskarazko gaitasunak automatikoki sailkatzeko sistema bat garatzeko eginkizuna burutzen saiatu gara.

Idatzizko euskara mailaren sailkatzaile bat sortzeko, geroz eta aukera gehiago daude. Euskaltegietan, hizkuntza-eskoletan eta oro har hezkuntza alorrean sortzen diren hizkuntza-probetako testuak formatu digitalean eskuratzeko eta datu horiek ikerketarako erabiltzeko aukera handiagotzen joan da azken urteetan. Honek, transkribatutako idazlan hauek erabilita, sailkapen sistemak garatzen hasteko lehen saiakerak egitea ahalbidetzen du.

Hizkuntza mailaren sailkatzaile eta ebaluatzaileak garatuta daude eta gero eta hobeak dira datu ugariko hizkuntzetarako, adibidez Senanayake eta Asanka (2024)-ek sistema bat garatu zuten errubrika eta testu bat emanda testuak ebaluatzeko. Euskaraz berriz, sistema ezberdinak sortzeko saiakerak egin dira, hainbat teknika erabiliz, Castro-Castro *et al.* (2008) edota Arrieta *et al.* (2023) adibidez. Horren adibide HiTZ taldeak garatutako lehen C1 sailkatzaile demoa [1] daukagu. Demo hau hobetzearren testu gehiago eskura izanda sailkatzaile hobeago bat

[1] https://huggingface.co/spaces/HiTZ/C1_sailkapen_demoa

garatzen saiatu gara.

Proiektu honetan garatuko den sistemaren erabilera praktiko bat euskara ikasten ari direnen komunikagaitasun-maila zehaztea da, ikasleen hizkuntza gaitasunak aztertuz. Hau baliogarria izango zaie ikasleei, irakasleei eta erakundeei; ikasleek haien maila jakiteko, irakasleek ikerketan edo ebaluazioan laguntzeko, eta erakundeek ebaluazio-neurri osagarri bat izateko.

Ikerketa honen helburua sarreran idazlan bat emanda C1 maila duen edo ez zehazten duen sailkatzaile bat garatzea da. Esperimentu ezberdinak garatu dira teknika ezberdinak erabilita ereduen errendimendua aztertzeko eta azkenik bi analisi egin dira ereduen portaera aztertzeko.

Zehazki ekarpen hauek egin ditugu:

- Euskarazko idatzizko C1 maila sailkatzen duen demo berri bat sortzea esperimentuetan lortutako eredu egokienarekin [2].
- Ataza honetarako lehen hurbilpena egitea euskaraz HE neuronalak erabilita, zehazki RoBERTa eta Latxa ereduak erabiliz.
- SCL, EDA eta erregularizazio teknikek duten eragina aztertzea.
- Ereduen artefaktuak eta portaera aztertzeko bi analisi mota diseinatu eta erabiltzea.

2 Arloko egoera eta ikerketaren helburuak

Testu sailkapenak testu edo hitz-sekuentziari aurrez definitutako etiketa bat esleitzea du helburu. Lengoaia Naturalaren Prozesamenduko (LNP) hainbat ataza testu sailkapen moduan ebazten dira, NLI, testu laburpen estraktiboa eta sentimendu analisia besteak beste. Tradizionalki Naïve Bayes Maron (1961) edota euskarri bektoredun makinak Joachims (1998) erabili ohi izan dira testu sailkapen sistemen oinarri gisa. Azken urteetan *transformers* arkitekturaren gorakadarekin HE kodetzaile eta sortzaileak erabiltzen dira nagusiki sailkapen atazak ebazteko. Azken hauek testuzko erantzunekin ebazteko aukera ere ematen dute, ereduaren arrazoiketa modu esplizituan erakutsi. Deskodetzaileak oso onak izanik ere kodetzaileak emaitza hobeak lortzen dituzte oraindik Etxaniz *et al.* (2024)-ek erakutsi zuten bezala.

Idazlanen puntuaketa automatikoa (IPA), ingelesez Automatic Essay Scoring (AES), oso ataza garrantzitsua eta ikertua izan da urte luzez (Page, 1967, 1968). Definizioz IPA atazak sarrerako testu bati aurrez definitutako puntuaketa sisteman zenbakizko nota bat esleitu behar dio. Hainbat hurbilpen eta datu-multzo ezberdin sortu dira ataza hau burutzeko, izan ere, idazlanak ebaluatzeko hainbat irizpide ezberdin erabili dira, narrazio kalitatea edota ideien antolaketa besteak beste (Kumar *et al.*, 2021). Page (1966)-ek lehen hurbilpena hitz kontaketa eta ezaugarri oso sinpleak erabilita egin zuen, gaur egun ordea ikasketa sakoneko teknikak erabiltzen dira ataza honetarako. Taghipour eta Ng (2016)-ek RNN eta sare konboluzionaletan oinarritutako hurbilpenak egin zituzten eta azken saiakerak HE sortzaileak erabilita egin dira (Stahl *et al.*, 2024).

Testuak Hizkuntzen Erreferentzia Marko Bateratuarekin (HEMB) lerrokatzea IPA atazen azpimultzo bat da, non testuaren puntuazioak eta errubrikak HEMBkoak diren. Hala ere, ataza ez dago oso zehatz definituta eta urteetan zehar hainbat aldaera sortu dira. Ikerketa batzuek testuak maila bat gainditzen duen edo ez zehazten dute, beste batzuk orokorrean A1-C2 tarteko maila bat esleitzen die.

Schmalz eta Brutti (2021)-ek BERT ereduak erabilita HEMB maila zehazten zuten EFCAMDAT Geertzen *et al.* (2013) eta CLC FCE Yannakoudakis *et al.* (2011) datu-multzoetan oinarrituta. Hurbilpen ezberdinak erabilita ere, ataza hau ez dago gaindituta, datu multzo artean ezberdin anotatzen direlako testuak eta ataza beraren definizio zehatzik ere ez dagoelako. Bestalde, ataza oso konplexua da hizkuntzaren ulermen gaitasun sakona behar baita eta datu eskasi eta anotazio marko bateraturik ez dagoenez sistemen garapenean zailtasunak daude.

Euskararako berriz hainbat hizkuntza mailaren sailkatzaile eta idazlanen puntuazio automatikoko sistemak garatu dira. Erregela, ezaugarri linguistiko eta ikasketa automatikoko teknika tradizionalak erabilita egin ziren hurbilpen hauek. Castro-Castro *et al.* (2008), Zipitria *et al.* (2010) eta Zipitria *et al.* (2011) edota Olaizola (2022)-en ikerketetan ezaugarri linguistikoak eta erregelak erabiltzen zituzten testuaren errepresentazio bektoreak lortu eta ondoren ikasketa automatikoko teknika tradizionalak erabili zituzten testuak ebaluatzeko. 2023an Arrieta *et al.* (2023)-ek HABE-IXA hitzarmenaren bitartez eskuratutako testuak erabilita sailkatzaile sistema bat garatu zuten.

[2] https://huggingface.co/spaces/HiTZ/C1_sailkapen_demoa1.1

Azpimultzoa	Kopurua	EZ_GAI %	GAI %
Hobe_Esk	614	%60.0	%40.0
22Api	4794	%63.8	%36.2
22Urr	5202	%70.3	%29.7

1. Taula: Azterketa epealdi bakoitzeko etiketen distribuzioa

Hurbilpen hau ezaugarri linguistikoak eta Euskarri Bektoredun Makinak (EBM) erabilita egin zuten. 4 sailkapen ataza definitu zituzten, lehena testuak dagokion maila gainditzen duen edo ez, bigarrena, B1-etik C2-rainoko maila zehaztea, hirugarren testuak B1-B2 edo C1-C2 mailetan dauden zehaztea eta azkenik, HABEren irizpide bakoitzeko nota zehaztea. Emaitza oso interesgarriak lortu zituzten, baina datu eskasia handia dela azpimarratu zuten.

2.1 Ikerketaren helburuak

Ikerketa honen helburua C1 mailako sailkatzailea garatzea da, hau da, sistemak sarreran idazlan bat jaso eta C1 gaindituko duen edo ez aurresan behar du. Horretarako, euskararako dauden HE eta teknologia aurreratuak ikertu dira.

Lan honetan hainbat ikerketa galdera definitu dira. Lehena, eskura dauden HE ezberdinekin esperimentuak egitea da, kodetzaileak eta deskodetzaileak erabili dira errendimenduak aztertzeko.

Behin HE egokienak identifikatuta ereduen errendimendua hobetzeko teknikak esploratu dira egokienak zein diren aurkitzeko. Jarraiko 3 teknikak erabili dira, xehetasun gehiago 3.2 atalean:

Erregularizazio teknikak aztertu dira izan gaindoitzeak eta alborapenak gainditzeko.

Supervised Contrastive Learning teknika esploratu da mugan dauden testuak egokiago sailkatzeko helburuarekin.

Easy Data Augmentation algoritmoa erabili da datu eskasiari aurre egiteko teknika gisa.

Azkenik, sortutako ereduen portaeraren analisiak egin dira, nota ezberdinetako testuekin ereduek duten zehaztasuna eta ereduek ikasitako artefaktuak aztertzeko.

3 Ikerketaren muina

Atal honetan ikerketa burutzeko erabilitako datuak, burututako esperimentuak eta ereduen portaeraren analisiak azalduko dira.

3.1 Datuak

Ikerketa honetan erabilitako testuak HABE-HiTZ arteko hitzarmen baten ondorioz eskuratutakoak dira. Zehazki idazlan bakoitzeko transkripzioa, HABEko ebaluatzaileek zehaztutako nota eta C1 maila gainditu duen (GAI) edo ez (EZ_GAI) daukagu. Bi azpimultzotan dago datu-multzo hau, transkripzioen iturrien arabera. Lehena *HOBE* izenez deitu dioguna eskuz transkribatutako testuek osatzen dute eta azterketa epealdi ezberdinetakoak dira. Bigarrena automatikoki transkribatutako testuak dira, bi epealditakoak *22Api* eta *22Urr*, hauetatik dugu testu gehien. Transkripzio automatikoen kalitatea gure ataza burutzeko nahikotzat jo da, bataz beste transkripzioen konfiantza maila %98 ingurukoa baita eta karaktere errore-tasa 2.20 ingurukoa. 1 taulak azpimultzo bakoitzaren testu kopurua eta klase banaketa erakusten dizkigu. Taulan ikusten den bezala EZ_GAI/GAI distribuzio banaketa nahiko orekatua da datu iturrien artean.

Azterketaldi batetik bestera idatzi beharreko testuen irizpideak eta formak mantentzen badira ere, idazlanen gaiak aldatu egiten dira. Gai ezberdinek sistemetan eragina dutela ikusi dugu ingelesez egindako aurreko esperimentuetan, entrenamenduan ikusitako gaiak errazago sailkatzen dituzte sistemek eta gaiez aldatzean errendimendua okertzen da. Gure idazlanak epealdi ezberdinetakoak direnez, testuen gaien eragina aztertzeko esperimentuetarako bi datu-multzo aldaera erabili ditugu: *Epealdika* entrenamenduan epealdi bakarra dago eta *Nahastuta* entrenamenduan bi epealdi nahasten dira. 2 taulak erakusten dizkigun entrenamendu/garapen/test banaketak sortu

Banaketa	Entrenamendua	Garapen	Testa
Nahastuta	4000 22Api + 2000 22Urr	2000 22Urr	614 Hobe Esk
Epeka	4000 22Api	2000 22Urr	614 Hobe_Esk

2. Taula: Esperimentuetarako egindako bi datu-multzo aldaeren banaketak.

ditugu epealdiak kontuan hartuta, bi aldaeren garapen eta test zatiak berdinak dira esperimentuak konparagarri izateko. GAI/EZ_GAI proportzioen oreka mantendu da.

3.2 Esperimentuak

Sekzio honetan ikerketan egin diren esperimentuak azalduko dira.

Oinarri lerroa Lehenik ondoren erabiliko diren teknika ezberdinen egokitasuna eta onura neurtzeko oinarri lerro batzuk entrenatu ditugu. Oinarrizko HEak aukeratzeko aurretiazko proba txikiak egin dira eta RoBERTa-euscrawl-large Artetxe *et al.* (2022) eredua hobea dela ikusi dugu Agerri *et al.* (2020)-en BertEus edota XLM-RoBERTa-rekin Conneau *et al.* (2019) alderatuta. Latxa HE sortzailea ere erabili dugu esperimentuak egiteko, baina testu sortzaile beharrean kodetzaile gisa erabilita. Oinarri lerroak bi datu-multzo aldaerak erabilita entrenatu dira, testuen gaiek izan ditzaketen eragina sakonago aztertzeko.

Erregularizazioa Behin oinarri lerroak izanda gaindoitzeak ekiditeko erregularizazio teknikak landu ditugu. Ereduak erregularizatzeko diluzio (*drop-out* ingelesez) probabilitatea igo diegu ereduei, RoBERTa ereduek 0,1 eta Latxak 0,0 erabiltzen dute defektuz. Diluzio probabilitatea igota ereduek orokortzeko duten gaitasuna areagotzea espero da, artefaktuak ikastea ekidinez. Artefaktuak ereduak ikasi dituen eta ataza burutzeko beharrezkoak ez diren ezaugarriak dira, hauek orokortzeko gaitasuna kentzen diote ereduari, ez baititu garrantzizko ezaugarriak ikasten.

RoBERTa ereduetan 0,3 balioarekin egin dira probak eta Latxa ereduetan 0,3 eta 0,4 erabilita probatu dira.

Easy Data Augmentation (EDA) (Wei eta Zou, 2019) datu eskasiari aurre egiteko lau metodo erraz eta merkeetan oinarritzen da, jatorrizko testuetatik datu sintetikoak sortzeko. Metodo hauek WordNet erabiltzen dute. EDAk memorizazio eta baliabide gutxiko eszenatokietan entrenatzean sor daitezkeen arazoak saihesten lagun dezake. Jatorrizko artikuluan, egileek teknika hauek erabiliz, baliabide gutxi eta ertainak dituzten datu-multzoetan errendimendua hobetzen dela erakusten dute. Gure atazan datu eskasia dugunez EDA bezalako teknikak lagungarri izan daitezke. EDAren heuristiko batzuk esaldia transformatu dezakete eta euskararen idatzizko maila okertu. Eragin negatibo hau neurtzeko, bi modutara probatu dugu EDA, lehenik EDAren lau heuristikoak erabilita eta bigarren sinonimoen ordezkapena bakarrik erabilita.

Heuristiko bakoitza esaldi mailan erabili da eta ondoren esaldiak lotu dira jatorrizko testuaren aldaera sortzeko. Esperimentuak sinplifikatzearren EDAren heuristiko bakoitzari 0,2ko ratioa ezarri zaio eta testu bakoitzaren 7 testu sintetiko sortu dira, guztira jatorrizko datu-multzoaren tamaina 8 aldiz handiagotuz.

Supervised Contrastive Learning (SCL) (Khosla *et al.*, 2020) jatorriz irudiak sailkatzeko teknika gisa diseinatu zen. Teknika honen intuizioa irudietatik lortzen diren errepresentazioek klase etiketaren arabera espazioan klusterrak osatzea da. SCL LNP arloan ere erabilgarria dela erakutsi dute ikerketa batzuk (Gunel *et al.*, 2020). Proiektu honetan mugan dauden idazlanak zailagoak izango dira muturrekoak baino sailkatzeko. Ereduak GAI edo EZ_GAI ezberdintzen errazago ikasteko SCL erabili dugu, mugako testuak errazago sailkatzeko.

Esperimentu hauetan SCL eta Cross-Entropy galera funtzioak aldi berean optimizatu dira. Bi galera balioen bataz bestekoa egin da konbinatzeko eta SCLren τ parametroa 0,2 balioarekin zehaztu da.

3.3 Esperimentuen emaitzak

Atal honetan 3.2 atalean deskribatutako esperimentuen emaitzak deskribatuko dira.

Ereduen errendimendua neurtzeko ohiko asmatze-tasaz (AT) gain idazlanek lortutako notaren arabera 3 multzo osatu dira eta multzo bakoitzeko AT kalkulatu da, horrela idazlan mota ezberdinekin ereduak duen portaera ebaluatuko da. Azterketen nota 0-30 tartekoa denez, [0-14), [14-16] eta (16-30] multzoak osatu ditugu eta bakoitzaren ATen izenak hauek dira hurrenez hurren, AT Baxua, AT Duda eta AT Altua. Gainditzeko muga 15 denez, AT Baxuak EZ_GAI diren testuak barne hartzen ditu, AT Altuak GAI direnak eta AT Duda mugan daudenak.

Banaketa	Eredua	AT Baxua	AT Duda	AT Altua	AT
Epeka	Majority Class	1,00	0,09	0,00	0,60
	RoBERTa-euscrawl	0,75	0,59	0,81	0,75
	+ EDA	0,77	0,45	0,67	0,70
	+ EDA-SO	0,74	0,61	0,77	0,73
	+ SCL	0,72	**0,64**	**0,83**	0,75
	+ diluzio-p 0.3	0,78	0,50	0,80	0,75
	Latxa	0,92	0,20	0,36	0,67
	+ diluzio-p 0.3	**0,83**	0,44	0,72	0,75
	+ diluzio-p 0.4	0,80	0,50	0,75	**0,76**
Nahastuta	Majority Class	1,00	0,09	0,00	0,60
	RoBERTa-euscrawl	0,84	0,45	0,73	0,77
	+ EDA	0.73	0.45	0.69	0.69
	+ EDA-SO	0,76	0,63	0,74	0,74
	+ SCL	0,71	**0,66**	**0,87**	0,75
	+ diluzio-p 0.3	0.78	0.63	0.81	0.77
	Latxa	**0,95**	0,23	0,36	0,69
	+ diluzio-p 0.3	0,86	0,45	0,72	0,77
	+ diluzio-p 0.4	0,85	0,47	0,77	**0,79**

3. Taula: Test ataleko emaitzak. Majority Class-ek gehien agertzen den klasea itzultzen duen eredua adierazten du, kasu honetan EZ_GAI.

Oinarri lerroen emaitzak 3 taulan ikus dezakegun bezala, Latxa ereduak 0,69-ko ATrekin ez dira RoBERTa ereduen 0,75 eta 0,77-eko ATetara gerturatzen. Epeka eta Nahastuta konparatzean, AT orokorrean hobea da Nahastuta entrenatzea, baina zehatzago ikusita AT Duda eta AT Altua hobea lortzen du Epeka entrenatutako ereduak. AT Baxua oso altu mantentzen da bietan, 0,80 ingurukoa izanik.

Erregularizazioaren emaitzak aztertuta orokorrean RoBERTa ereduetan diluzio gehiago erabiltzeak ez du onurarik ekarri. Hala ere, AT xeheagoak ikusita portaera aldatu dela nabaritu daiteke, Nahastuta aldaeran entrenatuta AT Duda eta AT Altuak hobekuntza nabaria daukate 0,18 eta 0,08 igo dira. Latxa ereduetan berriz, hobekuntza oso nabaria da. AT orokorrak 0,07-0.10 puntu igo baitira, diluzio probabilitate eta datu-multzo aldaeraren arabera. Diluzio gehiago erabilita emaitzek hobetzeko joera erakutsi dute. Erregularizatutako Latxa ereduak RoBERTak baina hobeak dira AT orokorrean, baina, RoBERTak oraindik oro har hobeak dira AT Duda eta AT Altuan.

EDA emaitzak ikusita teknika hau aplikatuta oinarri lerroak baina okerragoak diren ereduak lortzen ditugula ikus dezakegu. EDAko 4 heuristikoak aplikatuta emaitza okerragoak lortzen dira SO bakarrik erabilita baino. AT zehatzagoetan ere antzeko joera ikusten da, emaitzak oro har okerragoak dira.

SCL emaitzak orokorrean oinarri lerroarekiko berdinak dira. Hala ere, eredu hauek puntu batzuk gehiago lortzen dituzte metriketan Epeka entrenatuta eta Nahastuta banaketan aldiz hobekuntza nabariagoa da. AT Duda eta AT Altuaren balioak altuagoak dira SCL erabilita, baina AT Baxuan emaitzak apur bat okertzen direnez AT orokorra berdin mantentzen da.

3.4 Ereduen analisia

Esperimentuetako emaitzak aztertuta orokorrean badirudi eredu ia guztien portaera oso antzekoa dela, asmatzetasan gorabehera gutxi egonik. Eredu batzuek joera handia dute beti EZ_GAI esateko (AT Baxua altua dutenak), beste batzuk orekatuagoak dira. Domeinuaren eragina ere oso txikia dela dirudi Nahastuta eta Epeka entrenatuta emaitzak oso antzekoak dira eta. Ereduen arteko portaera sakonago aztertzeko bi analisi mota diseinatu ditugu: kalibrazio neurketa eta artefaktuen eragina.

Kalibrazio neurketak ereduek GAI esatean klaseari esleitzen dion probabilitatea eta testuaren notaren arteko korrelazioa neurtzen du. Zenbat eta idazlana 30etik gertuago egon probabilitate gehiago esleitu beharko lioke ereduak klaseari, 15etik gertu dagoen batekin zalantzatiago egon beharko luke eta %50-eko probabilitatea esleitu beharko lioke. Korrelazio hau neurtzeko pearson korrelazioa erabili da eta 3.3 atalean AT altuena atera duten

ereduak test zatian ebaluatu dira. 4. taulak erakusten digunez oinarri lerroko RoBERTa eta erregularizatutako Latxa ereduak dira korrelazio sendoena dutenak eta bereziki Nahastuta entrenatutakoak. SCL teknikak korrelazioan eragin negatiboa izan du eta EDA-SO-ren kasuan korrelazioa 0 ingurukoa da.

Banaketa	Eredua	Pearson korrelazioa
Epeka	RoBERTa-euscrawl	0.5502
	+ SCL	0.5394
	Latxa + diluzio-p 0.4	**0.6385**
Nahastuta	RoBERTa-euscrawl	0.6709
	+ EDA-SO	-0.0695
	+ SCL	0.5690
	Latxa + diluzio-p 0.4	**0.6713**

4. Taula: Ereduen pearson korrelazioa, balio altuagoak hobeak dira.

Artefaktuen eragina neurtzeko algoritmo sinple bat diseinatu dugu. Idazlan baten esaldiak ausaz berrordenatzean idazlanaren kalitatea asko jaisten da, GAI diren testuak EZ_GAI bihurtuz. Artefaktuak ikasi dituen eredu batek berrordenatutako testuak GAI direla aurresaten jarraituko du. Artefaktuekiko ereduek duten eragina aztertzeko, ereduek berrordenatutako esaldiak EZ_GAI direla zehazteko duten joera neurtu da test zatia ebaluatzeko. Sistema perfektu batek berrordenatutako testu guztiak EZ_GAI bezala ezarriko lituzke, metrikan 1,0-ko balioa lortuz. 5 taulan metrika honen balioak ikus ditzakegu ereduka. Latxa izan da eredurik egokiena eta RoBERTa ereduak atzean geratu dira gutxienez 10 puntuko aldea izanik, ondorioz, Latxa ereduak direla artefaktu gutxien ikasi dituztenak ondoriozta daiteke, bereziki Epeka entrenatutako eredua.

Banaketa	Eredua	Metrika
Epeka	RoBERTa-euscrawl	0.1176
	+ SCL	0.1856
	Latxa + diluzio-p 0.4	**0.2791**
Nahastuta	RoBERTa-euscrawl	0.0854
	+ EDA-SO	0.0072
	+ SCL	0.1610
	Latxa + diluzio-p 0.4	**0.2356**

5. Taula: Eredu bakoitzak berrordenatutako testuak EZ_GAI jartzeko duten joera, balio altuagoak hobeak dira.

4 Ondorioak

Ikerketa honetan idazlanek C1 maila gainditzen duten edo ez zehazten duen sistema bat garatu da. Publikoki erabilgarri jarri dugu RoBERTa eredu egokiena edonork proba dezan, Latxa ereduak ezin izan ditugu erabilgarri jarri konputazio kostu handiegia dutelako. Ereduak entrenatzeko bi datu-multzo aldaera probatu dira azterketaldi epeen arabera, Epeka, entrenamenduan epe bakarra duena eta Nahastuta, entrenamenduan bi epealdi nahastuta dituena. Oinarri lerroak hobetzeko 3 teknika ere probatu dira, erregularizazioa, EDA eta SCL. AT orokor altuena duen eredua Nahastuta datuekin entrenatutako eta erregularizatutako Latxa eredua da, 0,79-ko AT izanik.

3.3 atalean azaldu den bezala, erregularizatutako Latxa ereduak dira AT altuenak dituztenak, baina RoBERTa ereduak errendimendu orekatuagoa dute nota ezberdinetako idazlanekin. Hobekuntza teknikek ez dute onura esanguratsurik eragin ereduen errendimenduan, Latxaren kasuan erregularizazioak izan ezik.

Bi datu-multzo aldaeretan entrenatutako ereduen arteko aldea puntu gutxiko da eta kasu batzuetan Epeka egokiagoa da eta beste batzuetan Nahastuta, ez dago patroi argirik.

3.4 ataleko bi analisiek Latxa ereduak idazlanen notekiko korrelazio hobea eta artefaktuekiko eragin gutxiago dutela erakutsi dute.

5 Etorkizunerako planteatzen den norabidea

4 atalean deskribatu den bezala, Latxa ereduak izan dira bai AT altuenak lortu dituztenak, baita analisietan portaera egokiena erakutsi dutenak. Aurrera begira Latxarekin gehiago esploratzeko bidea ikusten da. Ikerketa honetan Latxa kodetzaile bezala erabili da eta ez dira bere sortzaile gaitasunak esploratu.

Artikulu honetan idazlanak modu bitarrean sailkatzea esploratu da eta ez da nota aurresateko biderik ikertu. Aurrera begira egokia izango litzateke erregresio sistemak garatzea eta sailkatzaileekin alderatzea.

Erreferentziak

Agerri, Rodrigo, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, eta Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*.

Arrieta, Ekain, Igor Odriozola Sustaeta, Xabier Arregi Iparragirre, eta Mikel Iruskieta Quintian. 2023. Habe-ixa euskarazko idazmen-proben corpuseko idazlanen mailakatze automatikoa. *Hizpide: helduen euskalduntzearen aldizkaria* 41.7.

Artetxe, Mikel, Itziar Aldabe, Rodrigo Agerri, Olatz Perez de Viñaspre, eta Aitor Soroa, 2022. Does corpus quality really matter for low-resource languages?

Azurmendi, Ekhi, eta Oier López de Lacalle. 2025. Euskarazko lehen c1 ebaluatzaile automatikoa. *arXiv preprint* .

Castro-Castro, Daniel, Rocío Lannes-Losada, Montse Maritxalar, Ianire Niebla, Celia Pérez-Marqués, Nancy C Álamo-Suárez, eta Aurora Pons-Porrata. 2008. A multilingual application for automated essay scoring. In *Advances in Artificial Intelligence–IBERAMIA 2008: 11th Ibero-American Conference on AI, Lisbon, Portugal, October 14-17, 2008. Proceedings 11*, 243–251. Springer.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, eta Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR* abs/1911.02116.

Etxaniz, Julen, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, eta Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. *arXiv preprint arXiv:2403.20266* .

Geertzen, Jeroen, Theodora Alexopoulou, Anna Korhonen, eta others. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, 240–254.

Gunel, Beliz, Jingfei Du, Alexis Conneau, eta Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403* .

Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, 137–142. Springer.

Khosla, Prannay, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, eta Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33.18661–18673.

Kumar, Rahul, Sandeep Mathias, Sriparna Saha, eta Pushpak Bhattacharyya. 2021. Many hands make light work: Using essay traits to automatically score essays. *arXiv preprint arXiv:2102.00781* .

Maron, Melvin Earl. 1961. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)* 8.404–417.

Olaizola, Xabier Azpillaga. 2022. Euskarazko testuen komunikagaitasun-maila automatikoki sailkatzeko lehendabiziko urratsak. *Hizpide: helduen euskalduntzearen aldizkaria* 40.3.

Page, Ellis B. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan* 47.238–243.

——. 1967. Grading essays by computer: Progress report. In *Proceedings of the invitational Conference on Testing Problems*.

——. 1968. The use of the computer in analyzing student essays. *International review of education* 210–225.

Schmalz, Veronica Juliana, eta Alessio Brutti. 2021. Automatic assessment of english cefr levels using bert embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics*.

Senanayake, Chamuditha, eta Dinesh Asanka. 2024. Rubric based automated short answer scoring using large language models (llms). In *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, volume 7, 1–6. IEEE.

Stahl, Maja, Leon Biermann, Andreas Nehring, eta Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845* .

Taghipour, Kaveh, eta Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1882–1891.

Wei, Jason, eta Kai Zou, 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

Yannakoudakis, Helen, Ted Briscoe, eta Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 180–189.

Zipitria, Iraide, Ana Arruarte, eta Jon A Elorriaga. 2010. Automatically grading the use of language in learner summaries. In *Proceedings of the 18th international conference on computers in education, Putrajaya, Malaysia*, 46–50.

——, Jon A Elorriaga, eta Ana Arruarte. 2011. Corpus-based performance analysis for automatically grading use of language in essays. In *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15*, 591–593. Springer.

6 Eskerrak eta oharrak

Artikulu hau master amaierako lan baten laburpen bat izan da, esperimentu, analisi eta xehetasun guztiak irakurtzeko jo jatorrizko dokumentura Azurmendi eta López de Lacalle (2025).

Proiektu honek honako finantziazioa izan du:

- IKER-GAITU proiektua 11:4711:23:410:23/0808 Eusko Jaurlaritza / Gobierno Vasco
- Eusko Jaurlaritzako Doktoretza-aurreko Programa PRE_2024_1_0035 espedientearekin