ORIGINAL

Gac Med Bilbao. 2024;121(2):62-68

Latxa-7b ereduan oinarritutako hizkuntzaren prozesamendu-sailkatzaileen gaitasunaren azterketa: medikuntzako aplikazioak eta kirurgia ortopediko eta traumatologiako testu klinikoen adibidea

Calvo-Lorenzo Isidoro[a]

(a) Servicio Vasco de Salud – Osakidetza. Organización Sanitaria Integrada Barrualde. Hospital Universitario Galdakao-Usansolo. Galdakao-Bizkaia.

Recibido el 16 de febrero de 2024; aceptado el 2 de abril de 2024

Giltza-hitzak

Hizkuntzaren Prozesamendua.
Ikasketa automatikoa.
Hizkuntza Eredu Handiak.
Kudeaketa Klinikoa.
Hezur-ehunen neoplasiak.

Laburpena:
Helburua:
Lan honetan Hitz Taldeko (Euskal Herriko Unibertsitatea) Latxa-7b Hizkuntza Eredu Handian egokitutako euskaraz idatzitako kirurgia ortopedikoko testu sintetikoen sailkatzaile baten gaitasuna aztertzen da.

Metodoak:
Datu-base sintetiko bat sortzen da, pazienteen 20.000 ohar klinikorekin, non patologia muskulueskeletikoen aipamenak agertzen diren. Latxa-7b-an oinarritutako sailkatzaile bat garatzen da, nota klinikoekin entrenatzen da eta hezurretako tumore gaiztoak detektatzeko duen errendimendua aztertzen da.

Emaitzak:
Sailkatzaile bat sortzen da, zeinaren errendimendua entrenamendu eta test datutaldetan % 97,7ko dointasunarekin, % 98,6ko zehaztasunarekin, % 94,2ko estaldurarekin, 0,99ko kurbaren azpiko eremuarekin eta 0,96ko F puntuazioarekin.

Ondorioak:
Lan honetan deskribatutako sailkatzailearen errendimendu bikainak akuilu izan beharko luke gure osasun-sistemetan erabiltzen ditugun historia kliniko digitalizatuetan hizkuntzaren prozesamendua aplikatzen hasteko.

Autor para correspondencia: isidro.calvolorenzo@osakidetza.eus (Calvo-Lorenzo Isidro).

Análisis de la capacidad de los clasificadores de procesamiento del lenguaje basados en el modelo latxa-7b: aplicaciones médicas y ejemplo de textos clínicos de cirugía ortopédica y traumatología

Resumen:
Objetivo:
En este trabajo se analiza la posibilidad de crear un clasificador de textos sintéticos de cirugía ortopédica escritos en euskera adaptado al Modelo de Lenguaje Grande Latxa-7b, creado por el Grupo Hitz (Universidad del País Vasco).

Métodos:
Se crea una base de datos sintética con 20.000 notas clínicas de pacientes en las que aparecen menciones a patologías musculoesqueléticas. Se desarrolla un clasificador basado en Latxa-7b; se entrena con notas clínicas y finalmente se analiza su rendimiento a la hora de detectar tumores óseos malignos.

Resultados:
Se crea un clasificador cuyo rendimiento en los grupos de datos de entrenamiento y test es de 97,7% de precisión, 98,6% exactitud, 94,2% de sensibilidad, 0,99 de área bajo curva y 0,96 de F1.

Conclusiones:
El excelente rendimiento del clasificador descrito en este trabajo debería servir de acicate para comenzar a aplicar el Procesamiento de Lenguaje Natural en las historias clínicas digitalizadas que utilizamos en nuestros sistemas sanitarios.

PALABRAS CLAVE

Procesamiento de Lenguaje Natural.
Aprendizaje Automático.
Modelo de Lenguaje.
Grande.
Gestión Clínica.
Neoplasias de Tejido Óseo.

Analysis of the capability of language processing classifiers based on the latxa-7b model: medical applications and example texts from orthopedic surgery and traumatology

Abstract:
Objective:
In this work we analyze the possibility of creating a classifier of synthetic orthopedic surgery texts written in Basque adapted to the Latxa-7b Large Language Model, created by the Hitz Group (University of the Basque Country).

Methods:
A synthetic database is created with 20,000 clinical notes of patients where there are mentions to musculoskeletal pathologies. A classifier based on Latxa-7b is developed. This classifier is later trained with clinical notes and finally its performance in detecting malignant bone tumors is analyzed.

Results:
A classifier is created whose performance in the training and test data sets is 97.7% precision, 98.6% accuracy, 94.2% sensitivity, 0.99 area under curve and 0.96 F1.

Conclusions:
The excellent performance of the classifier described in this work should serve as a spur to start applying Natural Language Processing to the digitized medical records we use in our healthcare systems.

Keywords

Natural Language Processing.
Machine Learning.
Large Language Model.
Clinical Management.
Bone Tissue Neoplasms.

Sarrera

Orain dela gutxi, Euskal Herriko Unibertsitateko Hitz taldeak Latxa euskarara bereziki egokitutako Hizkuntza Eredu Handiak (*Large Lenguage Models* ingelesez, LLM) oinarrizko ereduen bilduma sortu du. Eredu horiek Euscrawlekin euskal corpusarekin entrenatu ziren. 7.000 eta 70.000 milioi parametro bitarteko tartearekin, Latxa da gaur egun euskararentzat eraikitako LLM handiena eta errendimendu onenekoa[1].

Lan honetan Latxa-7b-an egokitutako euskaraz idatzitako kirurgia ortopedikoko testu sintetikoen sailkatzaile baten gaitasuna aztertzen da. Horretarako, hezurretako tumore gaizto gisa diagnostikatutako kasuen erreferentziak dituzten testu medikoak detektatzeko gaitasuna aztertuko da.

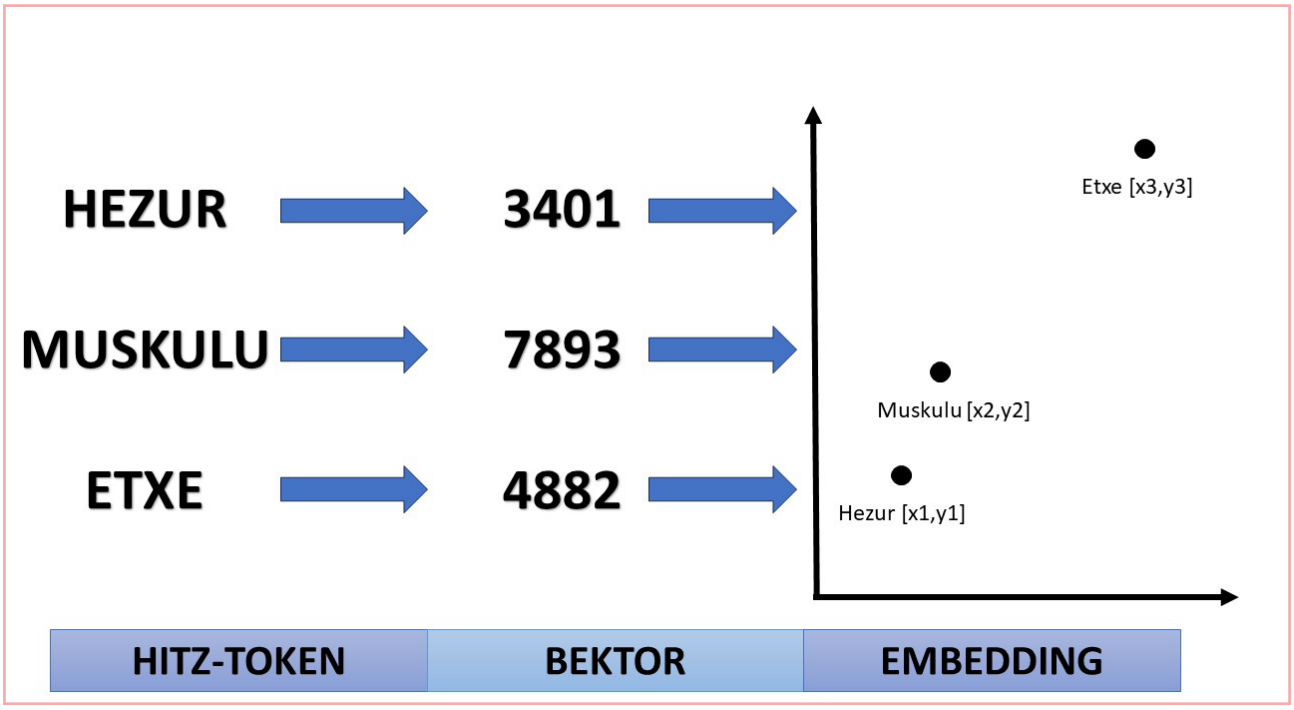
Era berean, Hizkuntzaren Prozesamenduaren (*Natural Language Processing* ingelesez, NPL) eta LLM-en funtzionamenduari buruzko argibide labur bat egiten da eta etorkizunean medikuntzan izango duten garrantzia azalduko da.

NPL, LLM-ak, eta beren garrantzia medikuntzako arloan

Sistema informatikoek lan egin dezakete informazio egituratuarekin (*structured data*), hau da, makinentzat ordena aurredefinitua eta ulergarria duten datuekin. Adibidez, Excel motako errenkaden eta zutabeen datubase batek datu egituratuak biltzen ditu. Hala ere, egungo makinak ez dira gai egituratu gabeko informazioa bereganatzeko, adibidez, Word-eko fitxategi batean gordetzen den testu librea. Gaur egun, osasun-erakunde gehienetan erabiltzen diren kudeaketa kliniko digitaleko sistemetan, gordetzen den informazioaren zati handi bat ez dago egituratuta[2]. Testu libreak, hala nola alta-txostenak edo eboluzio-txosten medikoak, irudi erradiologikoak, bideoak, elektrokardiogramak..., kudeaketa-sistema horietan gordeta daude, baina gizakiek bakarrik eskura ditzakete, makina batek ezin baitezake informazio hori ulertu.

NPL-ri esker, ordenagailuek giza hizkuntzan idatzitako testuak irakur ditzakete eta, beraz, testuetatik informazioa atera eta egitaratu dezakete. Horrela, historia kliniko digitalizatuetatik datu ugari datu-biltegietan metatzen dira, ondoren ustiatzeko (txostenak, alertak, kasuen kudeaketa, *machine learning*...).

Makinak ez dira gai gizakiok erabiltzen ditugun hitzek zer esan nahi duten ulertzeko. Ez dute inoiz asmatuko zer den etxe bat, mahai bat, hezur bat, tronbosi



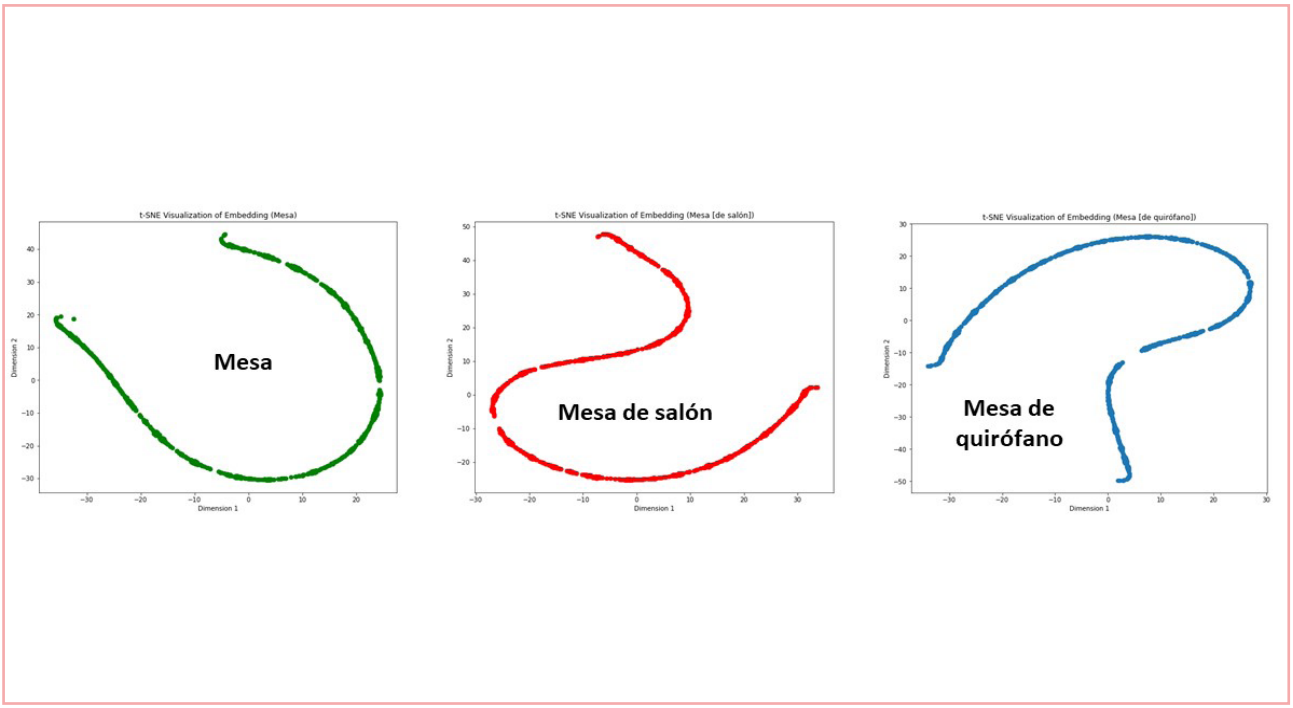
1. Irudia

benoso sakon bat... Zenbakiak bakarrik erabiltzeko gai dira. Horregatik, NPL-aren lehen urratsa hitzak testu-unitate txiki (tokens) bihurtzea da, ondoren zenbakizko balio bat esleitzeko (bektorizazioa). Horrela, konputagarriak ez diren sinbolo batzuk (hitzak) zenbakizko bektore konputagarri bihurtzen dira. Hala ere, makinak oraindik ez dira gai zenbakizko balio horrek zer adierazten duen ulertzeko. Beharrezkoa da bektore-token-

hitz hori beste ehunka mila bektore-token-hitzekin erlazionatzea. Horretarako, bektore bakoitzari koordenatu batzuk ematen zaizkio espazio kartesiar batean. Adibidez (1. irudia), "hezur" bektore-token-hitzak koordenatu jakin batzuk izango ditu (x_1, y_1). Koordenatu horretatik "muskulu" hitzera arteko distantzia (x_2, y_2) "etxe" hitzera baino txikiagoa izango da (x_3, y_3); beraz, makinak jakingo du "hezur" eta "muskulu" hitzek harreman handiagoa (distantzia txikiagoa) dutela bien artean "etxe" hitzarekin baino. Koordenatuak hiperplano batean esleitzeko prozesu horri (gizakiok erabiltzen ditugun hiru dimentsioak baino gehiago izan ditzake) embedding deritzo.

Embedding-ari esker, makinek hitzen arteko distantziak kalkula ditzakete eta, horrela, distanzia horien arteko erlazioak sor ditzakete. Kalkulu hauek dira NLP-aren oinarria. Duela urte gutxi arte, NLP sistema modernoak (NLTK, Spacy...) gramatikari bikainak ziren, eta hitz bakoitza oso zehatz lantzen zuten. Hala ere, hitzaren testuingurua galtzen zuten oso esaldi luzeetan. Horregatik, adibidez, itzultzaileek huts egiten zuten testu edo esaldi luze samarrak itzultzeko eskatzen zitzaienean. Egera hau arras aldatu zen transformer algoritmoaren agertzearekin 2017an 3. Transformer algoritmoari esker, makinek datu-kopuru erraldoiak erabil ditzateke Hizkuntza Eredu Handiak sortzeko. LLM hauen bidez, NPL-aren arloan iraultza bat gertatu da.

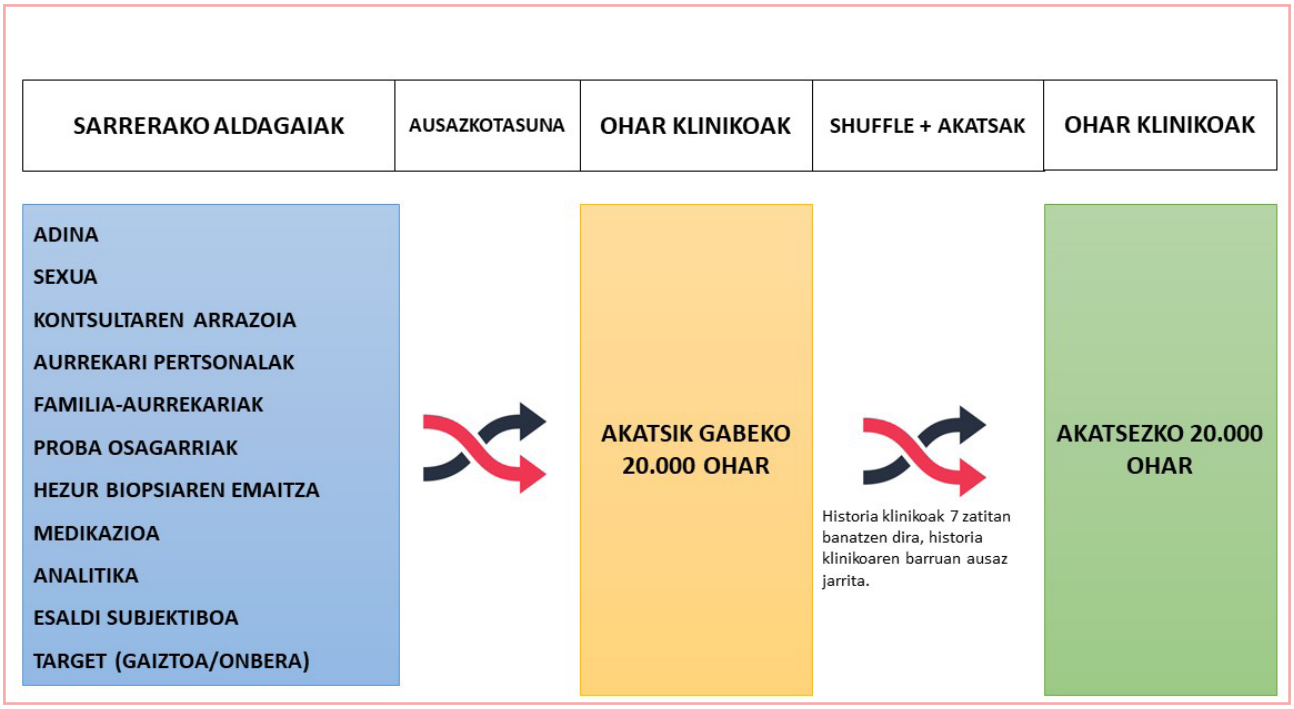
Transformer-a attention kontzeptuan oinarritzen da. Hau kalkulu algoritmikoa da, non hitz bakoitzaren esanahia esaldiaren inguruko hitzen arabera eraldatzen den. Hori lortzeko, bektore-token-hitz bati esleitutako koordenatuak ez dira finkoak, baizik eta esaldian duten



2. Irudia

posizioaren arabera eta erlazionatzen diren hitzen arabera aldatzen dira (2. irudia). Transformer bat entrenatzeko, datu-bolumen erraldoiak eta konputazio-gaitasun handiak behar dira, eta hori enpresa teknologiko handien edo ikerketa-taldeen esku dago soilik. LLM bat datu-bolumen erraldoiekin aurreentrenatutako transfomer algoritmo bat izango litzateke. LLM ugari daude argitaratuta, hala nola BERT (Google), LLaMA (Meta), GPT (Openai), edo lan hau jorratzen duena, Latxa (Hitz Taldea). Printzipioz, LLM batek ez du inolako zereginik egiteko gaitasunik: beharrezkoa da, aldez aurretik, LLM horrek entrenamendu bat jasotzea, baina datu-bolumen eta konputazio-gaitasun askoz txikiagoarekin LLM bera garatzeko behar dena baino. Horrela, LLM-etatik abiatuta, itzultzaileak, sailkatzaileak, informazio-erauzgailuak edo chatbotak bezalako tresnak entrenatu eta sor daitezke. Adibidez, ChatGPT chatbot ospetsua LLM aurreentrenatu baten aplikazio praktikoa da (Generative Pretrained Transformer edo GPT).

LLM-ak posible egiten duen teknologia klinikoko sistema digitalizatuetan aplikatzeak, pazienteak maneiatzeko oso datu baliotsuak berreskuratzeaz gain, pazientearen segurtasuna hobetuko luketen eta klinikoei eta ikertzaileei beharrezkoa ez den lan-karga



3. Irudia

murriztuko lieketen aplikazioak ezartzea ahalbidetuko luke[4].

Materiala eta ondorioak

Datu-base sintetikoa sortzea (3. irudia)

Patologia muskulueskeletikoak dituzten pazienteen 20.000 ohar klinikoko database bat sortzen da. Ohar kliniko bakoitzak 7 zati edo esaldi ditu:

1. zatia: adina (ausazkoa 18-99 tartean), sexua (ausazkoa: gizona edo emakumea), ukitutako hezurra (ausazkoa: tibia, peronea, femurra, humeroa, besaurrea) eta lateralitatea (ausazkoa: eskuina eta ezkerra)

2. zatia: kontsultaren arrazoia, aurrekari pertsonalak eta familiarrak

3. zatia: Pazienteak hartzen dituen medikazioak (77 printzipio aktibo desberdinen artean ausaz hautatutako 2-10 preskripzio)

4. zatia: Proba osagarriak

5. zatia: Biopsiaren diagnostikoa (ausazkoa: “gaiztoa”, “onbera” eta “gaiztoa izatea baztertzen da”). Komeni da sailkatzailea gaizto- edo onberatasun-kasuak detektatzeko gai izateaz gain, diagnostiko negatiboen logika ulertzea (adibidez, biopsia batek tumore gaiztoa baztertzen badu, onberatasun-kasu bat da).

6. zatia: Analitikaren emaitza. Maila batzuen barruan ausazko balioa ematen zaie glukosari (60-120), sodioari (130-150), kreatininari (0,5-1,2), PCRari (1-30), hemoglobinari (9-15) eta leukozitoei (5-15)

7. zatia: Pazientearen esaldi "subjektiboa"

Erabilitako esaldiak (kontsultarako arrazoia, aurrekari pertsonalak, aurrekari familiarrak, biopsiaren emaitza, froga osagarriak eta esaldi subjektiboa) esaldi-corpus luze baten barruan aukeratzen dira, ausaz, eta horietako batzuk sexuaren arabera bereizten dira (kontsultarako arrazoia, aurrekari pertsonalak eta proba osagarriak).

Zazpi zatiak ausaz ordenatzen dira, hots, 7! = 5040 modu desberdin daude ohar klinikoen informazioa antolatzeko.

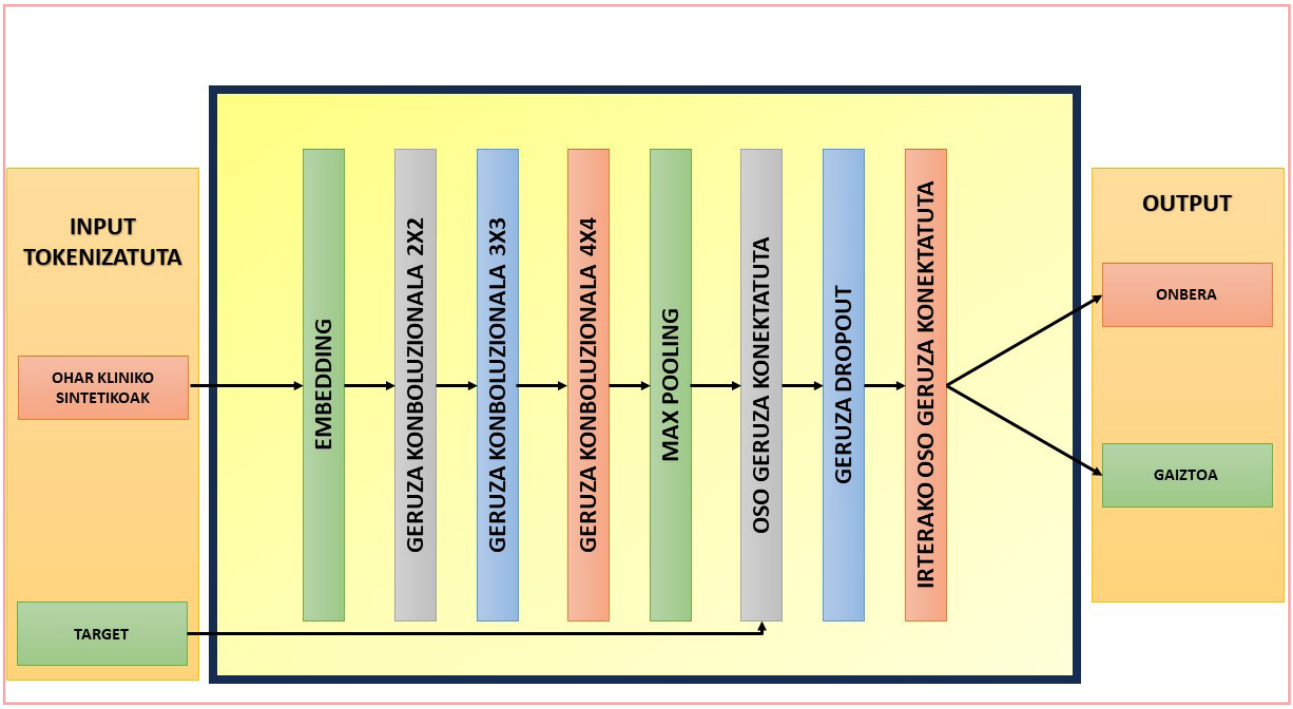
Azkenik, ortografia-akatsak sartzen dira ohar klinikoetan, % 1eko akats odds-arekin.

Horrela, 20.000 nota kliniko heterogeneoak lortzen dira, ortografia-akats ugarirekin. % 25 hezur-tumore gaiztoei dagozkie (sailkatzailearen helburua eto target-a), beste % 25 hezur-tumore onberak dira, eta % 25 muskulu-hezurretako beste patologia batzuk (azken finean, onberak ere bai). Azkenik, kasuen % 25etan patologia gaiztoa baztertzen da ("gaiztoa" terminoa aipatzen da, baina kasu onberak dira).

Era berean, sailkatzailearen azken testa egiteko, ortografia-akatsik gabeko 20.000 oharreko database bat gordetzen da. Kodea hemen dago eskuragarri: https://www.kaggle.com/code/isidorocalvolorenzo/latxa-esaldiak

Datuen prestaketa

Datuen koherentzia bermatzeko, testuak ondorengo prozesuak alda ditzaketen sinbolo guztiez garbitzen dira. Errendimendua optimizatzeko eta konputo aurrezteko, sailkatzailea garatzen hasiko da lehenengo 5000 ohar klinikoekin, eta errendimendu-emaitzen arabera handitu ahal izango da kopurua.



4. Irudia

Tokenizazioa, vektorizazioa eta embedding
Ohar klinikoak tokenizatzeko LLM Latxa-7b-a erabiliko da. Horretarako, LLM hau inplementatuko da Meta taldeko Llama2 tresnekin. Tokenizazioa ondo funtzionatzen duen egiaztatu ondoren, sailkatzailea sortzeko erabiliko diren 5.000 ohar klinikoak tokenizatu eta vektorizatu egiten dira. Hortik aurrera, embedding prozesua has daiteke, eta, lehen azaldu den bezala, testuen hitz bakoitza n dimentsioko hiperplano batean erlazionatzea ahalbidetuko du.

Sailkatzailearen definizioa
Sailkatzailea entrenatzeko, embedding egin eta gero, ohar klinikoak sare neuronal konboluzional baten bidez eraldatzen dira (deep convolutional neural networks, DCNN). DCNN sare hauek irudi digitalei machine learning teknikak aplikatzeko erabili ohi dira, baina testu klininoekin lan egiteko ere oso baliagarriak dira 5. Horretarako, 7 geruza (4. irudia) dituen sare neuronal bat diseinatzen da.

Sailkatzailearen entrenamendua eta ebaluazioa
Hautatutako 5000 nota klinikoak bi taldetan banatuko dira ausaz, bata entrenamendukoa (%90) eta bestea testekoa (%10). Entrenamendu-taldea aurrez deskribatutako sare neuronala entrenatzeko erabiliko da. Entrenamendu honetatik ohar klinikoen sailkatzaile bat lortuko da, hezurretako tumore gaiztoaren kasu positiboak ditu detektatzen dituena. Ondoren, sailkatzailea test taldeko ohar klinikoen % 10arekin probatuko da. Era berean, testu libreak probatuko dira, sailkatzaileak entrenatu denaz bestelako testuak sailkatzeko zer gaitasun duen baloratzeko.

Ereduaren azken ebaluazioa
Ortografia-hutsegiteak sortzeko kodea aplikatu ez zaien 20.000 ohar kliniko dituen datu-basetik abiatuta, 1000 ausaz hautatuko dira, eta lortutako sailkatzaileari aplikatuko zaizkio. Nahaste-matrize bat (confusion matrix) sortuko da, eta errendimenduaren analisisa egingo da.

Erabilitako tresnen deskribapena
Prozesu osoa Python programazio lengoaiaz egin da. Konputazio-gaitasunaren arazo lokalak direla eta, eta kontuan hartuta ez direla pazienteen datu errealak erabiltzen, sailkatzailea Kaggle plataforman garatu eta gauzatu da. Kodea hemen dago eskuragarri: https://www.kaggle.com/code/isidorocalvolorenzo/latxa-clasificator

Emaitzak
Ohar klinikoak prozesatzen dira, aurrez azaldutako metodologiaren arabera. Horrela, sailkatzaile bat sortzen da, zeinaren errendimendua entrenamendu-taldean % 97,7ko dointasunarekin (accuracy), % 98,6ko zehaztasunarekin (precision), % 94,2ko estaldurarekin (recall), 0,99ko kurbaren azpiko eremuarekin (ingelesez, Receiver Operating Characteristic curve, ROC) eta 0,96ko F puntuazioarekin. Sailkatzaileari test taldea aplikatzen zaionean, emaitzak berdin-berdinak dira. Emaitza horiek adierazten dute ez dagoela gehiegizko doikuntza-

I Taula
Sailkatzailaren errendimenduak, datu-multzoen arabera

	TRAINING DATA	TEST DATA	FINAL TEST DATA
Dointasuna	97,70%	97,70%	100%
Zehaztasuna	98,60%	98,60%	100%
Estaldura	94,20%	94,20%	100%
ROC	0,99	0,99	1
F puntuazioa	0,96	0,96	1

rik (*overfitting*), eta, beraz, sailkatzaileak ohar klinikoak ondo klasifikatzeko beharrezko ereduak ikasten lortu duela. (I taula)

Sailkatzaileari testu libreko zatiak aplikatzen zaizkionean, ereduak huts egiten du. Adibidez, biopsia negatiboak detektatzeko gai den arren ("Osteosarkoma baztertu daiteke" onbera bezala sailkatzen da), perpausaren logika konplexuagoa denean huts egiten du ("Ezin da tumore gaizto bat baztertu" onbera bezala sailkatzen da eta). Bestalde, logikoa denez, sailkatzaileari erakusi ez zaion informazioa erabiltzen denean, honek ezin du sailkapen zuzenik egin (sailkatu du "Gripea izan du" gaizto gisa) (5. irudia).

Azken testean, sailkatzaileak zuzen sailkatzen ditu ortografia-akatsik gabeko 1000 nota klinikoen % 100, eta errendimendu horrek gainditu egiten du entrenamendu-testuekin lortutakoa.

Eztabaida

Gaur egun, LLM-ak aldatzen ari dira gizakiek ordenagailuekin komunikatzeko daukaten modua. Teknologia honetan oinarritutako chatbot-en arrakasta, esaterako ChatGPT-arena, adibide argia da. Klinikari dagokionez, ordea, atzerapena dago teknologia hori historia klinikoetan baliatzeko orduan. Atzerapen horren arrazoia da, besteak beste, Adimen Artifizialeko inguruneetan

```
Ezin da tumore gaizto bat baztertu
Output: [[0.04199798]]
Aurreikusitatua: Onbera.


Osteosarkoma baztertu daiteke
Output: [[0.27894163]]
Aurreikusitatua: Onbera.


Tibia hautsi zaio
Output: [[0.8440923]]
Aurreikusitatua: Gaiztoa.


Gripea izan du
Output: [[0.7868743]]
Aurreikusitatua: Gaiztoa.


Biopsiak biriketako minbizia baztertzen du
Output: [[1.9761133e-05]]
Aurreikusitatua: Onbera.
```

5. Irudia

datu klinikoen erabilerak mugatuko eta zehaztuko dituen legeriarik ez dagoela[6].

Euskarak —beste hizkuntza batzuekin alderatuz gero[7], hala nola espainiera edo ingelesa— testu digitalen corpus oso mugatua dauka. Gainera, Euskal Autonomia Erkidegoko ospitaleetan erabiltzen diren testu kliniko gehienak gaztelaniaz idazten dira, eta, hortaz, euskarazko LLM-ak entrenatzeko gaitasuna oso mugatua da. Beraz, oso zaila da Adimen Artifizialean oinarritutako tresna klinikoak (hala nola, sailkatzaileak, informazio-berreskuratzaileak edo, are gehiago, chatbot-ak) sortzea hizkuntza horretan.

Hala ere, Hitz zentroak, Latxa LLM-ren sorreraren arduradunak, erakutsi du ez dela testuen corpus oso handirik behar etekin bikainak lortzeko[8]. Bestalde, ohar kliniko sintetikoak erabiliz lan honetan egin den probak erakutsi du, Latxa testu medikoekin berariaz aurreentrenatutako LLM bat ez bada ere, balio duela (adibidez, gaztelaniazko Roberta-bioklinikoa[9]), testu klinikoekin lan egiten duten erremintak sortzeko oinarri gisa.

Lan honek zenbait muga ditu: konputazio falta azpimarratu behar da, eta, horren ondorioz, ereduan entrenatzeko erabili diren ohar klinikoen kopurua arras txikia izan da. Bestalde, Latxa LLM arinena erabili da, hau da, Latxa 7b-a, eta ez indartsuena, 70b-a. Datuak eta parametroak murriztuta ere, emaitza bikainak lortu dira, errendimenduari dagokionez.

Beste muga bat erabilitako datu sintetikoen izaerarena izan da. Aldez aurretik, ChatGPT bezalako chatboten bidez historia klinikoen sorkuntza masiboa probatu zen gaztelaniaz, eta emaitzak nahikoa etsigarriak izan ziren[10]. Beraz, LLM-etan oinarritutako beste sailkatzaile kliniko batzuekin ikusi den moduan[11], Latxarekin ere giza-gainbegiraletza funtsezkoa da. Ausazko esaldiak erabiltzeak, ausaz ordenatzeak eta ortografia-akatsak sortzeak ez die homogeneotasun handiegia kentzen sortutako testuei, eta, hortaz, antzerako testu-egiturak erabiltzen direnean funtzionatzen du soilik sailkatzaileak. Hala ere, sailkatzaileak, entrenamenduan zehar, eskuragarri izan ez duen akatsik gabeko datu-basearekin lortutako errendimendua entrenamenduko datuekin baino hobea izan izanak, Latxa LLM-ak euskarazko testuekin lan egiteko daukan gaitasun bikaina berresten du.

Ondorioak

Nahiz eta aurrez deskribatutako mugak izan eta garapen-fasean dagoen eredua izan, badirudi Latxa LLM-ak teknologia horiek euskarazko testu medikoetan etorkizunean aplikatzeko aukera emango duela. Lan honetan deskribatu dugun sailkatzailearen oinarri gisa duen errendimendu bikainak akuilu izan beharko luke gure osasun-sistemetan erabiltzen ditugun historia kliniko digitalizatuetan hizkuntzaren prozesamendua aplikatzen hasteko.

Bibliografia

1. Model Card for Latxa 7b. HiTZ/latxa-7b-v1 · Hugging Face
2. Kong HJ. Managing Unstructured Big Data in Healthcare System. Healthc Inform Res. 2019; 25:1-2. doi: https://doi.org/10.4258/hir.2019.25.1.1
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. 2017:1706.03762. https://arxiv.org/pdf/1706.03762.pdf
4. Yang R, Fang Tan T, Lu W, Thirunavukarasu AJ, Wei Ting DS,Liu N. Large language models in health care: Development, applications, and challenges. Health Care Sci. 2023; 2: 255-263 doi: https://doi.org/10.1002/hcs2.61
5. Yao L., Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak. 2019; 19 (Suppl 3). doi: https://doi.org/10.1186/s12911-019-0781-4
6. Thirunavukarasu AJ, Ting DSJ., Elangovan K, Gutierrez L, Fang Tan T, Wei Ting DS. Large language models in medicine. Nat Med. 2023;29: 1930–1940. doi: https://doi.org/10.1038/s41591-023-02448-8
7. Agerri R, San Vicente I, Campos JA, Barrena A, Saralegi X, Soroa A, et al. Give your Text Representation Models some Love: the Case for Basque. Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 4781–4788.
8. Artetxe M, Aldabe I, Agerri R, Perez-de-Viñaspre O, Soroa A. Does Corpus Quality Really Matter for Low-Resource Languages?. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 7383–7390.
9. Solarte-Pabón O, Montenegro O, García-Barragán A, Torrente M, Provencio M, Menasalvas E, et al. Transformers for extracting breast cancer information from Spanish clinical narratives. Artificial Intelligence in Med. 2023; 143, 102625. doi: https://doi.org/10.1016/j.artmed.2023.102625
10. Calvo-Lorenzo I, Uriarte-Llano I. Generación masiva de historias clínicas sintéticas con ChatGPT: un ejemplo en fractura de cadera. Med Clin. 2024. Artikulua prentsan. doi: https://doi.org/10.1016/j.medcli.2023.11.027
11. Singhal K, Azizi S, Tu T, Mahdavi S, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023; 620: 172–180 (2023). Doi: https://doi.org/10.1038/s41586-023-06291-2