NER eta RE, sare soziala batean erabaki klinikoak hartzeko

Xabier Larrayoz
xlarrayoz001@ikasle.ehu.eus

Abstract

Hizkuntzaren prozesamenduaren zeregin nagusiak, hala nola entitateen erauzketa eta harremanen erauzketa, medikua bezalako domeinu espezifiko batean erronka irekia dago, eta are gehiago informazio-iturria sare sozialetatik badator, non hizkuntza informal prozesamendu egokia eragozten duen. Lan honetan erronka hori aurre egiteko hainbat irtenbide planteatzen dira. Horretarako, entitate eta harreman medikoekin tweetak ustiatu gabeko corpus batekin lan egingo da.

Bi zeregin horien ikuspegi bateratuarekin lan egingo da, eta entitateen kasuan, tarteka hurbilduko da. Egindako aurreprozesuarekin konbinatuta, F-1ean % 80 inguruko puntuazioak lortu dira zenbait entitateentzat. Horrek guztiak baliabide gutxiko ingurunean, etorkizunean eremu horretan lan egitearen ondorioz sortzeko aplikazioak aukera indartzen du.

1 Sarrera

Prozesatu beharreko informazio ugari sortzen den testuinguru batean, Transformersen oinarritutako aurre-entrenatutako hizkuntzaren ereduek, hala nola BERT, RoBERTa eta T5, aro berri baten gorakada markatu zuten hizkuntzaren prozesamenduan. Transformerrak, ikaskuntzaren transferentzia eta ikaskuntza auto-superbizituaren konbinazioaren ondorioz, azken urteetan azken hamarkadetan baino lorpen gehiago lortu dira. Milioika dokumentutatik informazioa atera ahal izatea arlo profesional askotan presentzia duen tresna indartsua da (Peng et al., 2019).

Medikuntzaren arloak, aspaldidanik, adimen artifizialean oinarritutako hainbat tresnaren laguntza du. Entitateak identifikatzea (named entity recognition, NER) eta haien harreman semantikoak (relation extraction, RE) funtsezko zereginak dira informazioa ateratzeko. Domeinu honetan, bai NERek eta bai REk gaixotasunak iragartzen, erabakiak hartzen eta pazienteen kohorteak identifikatzen lagun dezakete.

Azken hamarkadan, Twitter edo Reddit bezalako sare sozialek presentzia handiagoa izan dute pertsonen bizitzan. Erabiltzaileek bere pentsamenduak eta gogoetak adierazteko aukera lortzen dute, milaka pertsonarengana iristen. Gainera, informazio-iturri nagusietako bat bihurtu da.

Egunero, milioika erabiltzailek bere esperientzia partekatzen dute sendagai jakin batzuekin edo beraien egoera fisiko edota mentala. Pazientearen bilakaeraren etengabeko erregistroa, medikuarekin egindako kontsultan jasotzen ez dena.

Informazio-iturri batzuekin lan egin arren, batzuk alferrik galtzen dira, sare sozialen kasuan bezala. Erabiltzaileek sortutako testua denez, zailtasun bereziak ditu. Hizkuntza informalak zaildu egiten du behar bezala prozesatzea.

2 Aurrekariak

2018an, BERTk, Devlin et al. proposatutako ereduak, Wikipediaren eta google books-en edukiarekin aurre-entrenatuta, artearen egoera berri bat ezarri zuen hizkuntz naturalaren prozesamenduaren ataza gehienetan. Une horretatik aurrera, BERTren bertsio afinatuak sortu dira, ingurune desberdinetara egokitzeko (Li et al., 2021; Guo et al., 2021).

BERTk ikasten duen hizkuntzaren irudikapenak berak eragotzi egiten dio eremu espezifikoetara, bio-medikuntzara adibidez, ondo egokitzea. Sektore horietan erabilitako terminologiak ez du zuzeneko ordezkaritzarik ereduak ikasitakoan. Ondorioz, ereduen errendimenduak behera egin du, eremu orokorragoekin alderatuta.

Testu zientifikoen kopurua mugatuagoa denez, eremu horretan espezializatutako eredu bat garatzea zailtzen du. Eredu bat entrenatzeko, corpus hibrido bat erabiltzen da, domeinu corpusa domeinu orokorra edo erlazionatuta testu batekin

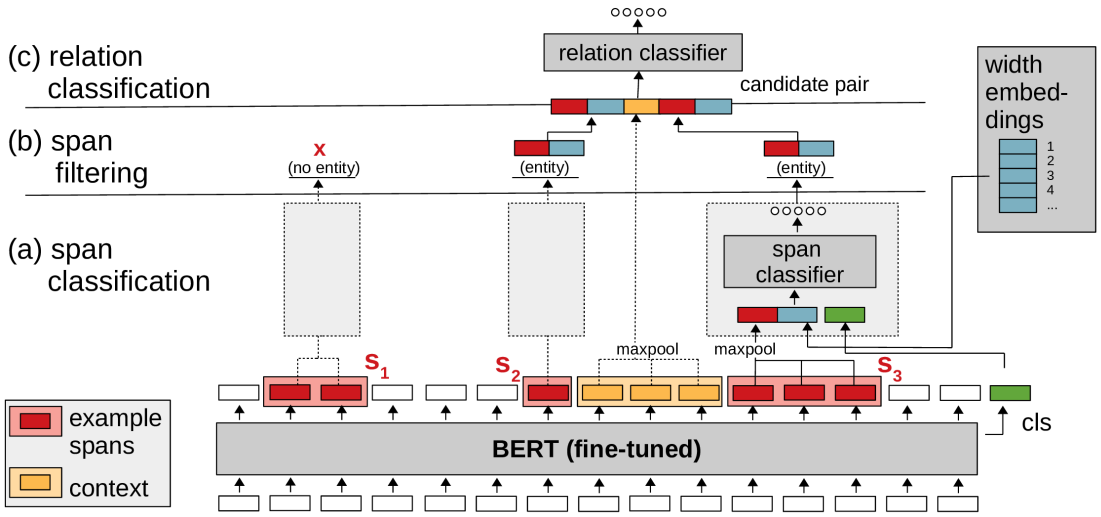

Figure 1: SpERTen arkitektura

konbinatuz. Hala ere, eredu bat hutsetik entrenatzeak dakarren denbora eta baliabide kostua dela eta, hainbat aukera aztertu dira. GreenBioBERTek BERT hedatzea planteatzen du, embeddingaren geruza finduz (Poerner et al., 2020). Bestalde, Tai et al. modulu gehigarri bat proposatzen dute exBERT ereduan, BERT embedding-a osatuko lukeen parametro ikasgarriekin.

Eremu biomedikoan, BLURB proban lortutako nota artearen egoeratzat hartuta, 6 zereginetan % 84,30 da batez beste (Yasunaga et al., 2022). Hala ere, erabilitako dataset guztiak oso urrun daude sare sozialetan lehen aipatutako propietateetatik. Bio-medikuntzaren domeinu espezifikoa eta sare sozialetatik lortutako zarata konbinatzen dituzten lanek % 74ko emaitzak lortzen dituzte F-1 mikro metrika erabiliz (Scepanovic et al., 2020; Foufi et al., 2019).

3 Sistema

Sarritan, NER eta RE atazak bereizita jorratu dira, mailakako ikuspegi batekin, non entitateen erauzketa harremanen aurretik dagoen. Horren ondorioz, akatsak sistema batetik bestera zabal daitezke, irteera okerrak sortuz eta irismena mugatuz. Multi-Task Fine-Tuning bat eginez, berriz, modeloa bi zereginetan entrenatzen da aldi berean. Horrela, balizko erroreak zabaltzea edota overfitting ematea saihestuko da, baliabide gutxiko agertokietan eraginkortasuna handitzen den bitartean (Giorgi et al., 2019; Kalyan et al., 2021).

Horiek horrela, SpERTren arkitektura erabiltzea planteatu da, NER eta RErako eredua, tarteetan oinarritua. Lanak hainbat doikuntza egiten ditu embedding-ean, emaitzak hobetzeko. Tarteen bidez hurbiltzen denez, gainjarritako erakundeak identifikatzeko gai da (Eberts and Ulges, 2019; Li et al., 2021).

1 irudian ikusten denez, aldez aurretik tokenizatutako sarrera entrenatutako modelo batetik pasatzen da, eta dagokion embedding-en sekuentzia lortzen da. Sor daitezkeen tarteak entitate ezagunen batean sailkatzen dira, sekuentziaren testuingurua erabiliz. Entitate bati ez dagozkion tarteak iragazi ondoren, erlazioen sailkapena egiten da, entitateen arteko testuingurua erabilita. Sarrera osoaren irudikapenaren ezagutza globala soilik erabiltzen duten beste lan batzuk ez bezala.

Esperimentu honetarako, BERT gain, propietate eta ingurune desberdinak dituzten hainbat eredu aztertu dira.

BioBERT, menderatze medikorako hizkuntza espezializatuaren eredu bat da, BERTtik abiatuz medikuntzari buruzko artikuluen corpus batean trebatua izan zen. Hizkuntza menderatzearen irudikapen espezifikoa lortuz (Lee et al., 2019).

BioRedditBERT, BioBERTetik abiatua, baina osasunaren sektorearekin lotutako Redditen foroetatik lortutako mezuekin entrenatua (Basaldella et al., 2020).

Bio-ClinicalBERT, BioBERTetik abiatuta, hainbat ospitaletako pazienteen erregistro medikoekin entrenatu zen, pazientearen zein familiaren gaixotasunen historialera sartzeko aukerarekin (Alsentzer et al., 2019).

CT-BERT, COVID-Twitter-BERT, domeinu medikoa sare sozialen testuinguru batean konbinatuz entrenatu zen beste eredu bat da. Kasu hone-

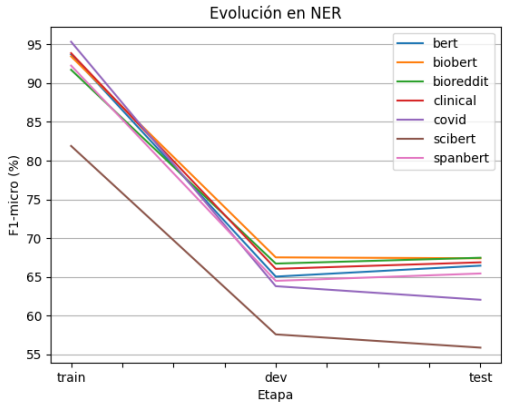

Figure 2: Ereduen NER atazan lortutako errendimenduak train,eval eta test zehar

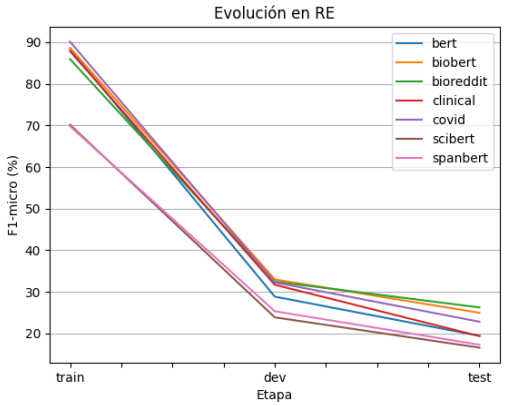

Figure 3: Ereduen RE atazan lortutako errendimenduak train,eval eta test zehar

tan, Covid-ari buruzko Twitterreko mezuak erabiliz (Müller et al., 2020).

SciBERT, paper zientifikoen corpus itzel batekin entrenatu zuten, eta bere hiztegi espezializatua sortu zuten (Beltagy et al., 2019).

SpanBERT, BERTren datu berberekin emaitza hobeak lortu zituen, tartekako entrenamendu-metodoari esker, tokenen segidako tarteak ezkutatzen ditu, ausazko tokenak ez bezala (Joshi et al., 2019).

Eredu guztiak 1 taulan adierazitako hiperparametro berberekin erabili dira.

epochs	15	size embedding	25
neg entity count	100	neg relation count	100
lr	5e-5	lr warmup	0.1
weight decay	0.01	max grad norm	1.0

Table 1: Erabilitako hiperparametroen konfigurazioa

4 Datuak

Ikerketa BEARi buruz egin da, Twitterretik berriki lortutako corpusari buruz (Wührl and Klinger, 2022). Bio-medikuntzari buruz hainbat erabiltzailek bidalitako mezuen bilduma da multzoa. Pazienteen ikuspegia eskainiz eta kontsulta batean bildutakoaz bestelako informazioa emanez. 2.100 tweeten artean, guztira 6.000 entitate idatzi ziren 14 klasetan banatuta eta 3.000 erlazio 20 klasetan. Interes berezikoa da, datu horien jatorriagatik ez ezik, erakunde eta erlazio mota ugari dituelako, beste data-set ez bezala, non entitate pare batzuk soilik dituztenik.

Datuekin lan egin ahal izateko, formatua SpERTrekin bateragarria den formatu batera egokitu behar izan zen. Testuaren aurreprozesua emotikonoak testu bihurtzea izan zen, eman zezaketen informazioa aprobetxatzeko. Zenbait alderdi normalizatzeaz gain.

Datu multzoaren izaera dela eta, hainbat erakunde eta harreman bateratu ziren ordezkaritza hobea lortzeko. Ondorioz, ehuneko-puntu batzuk hobetu ziren emaitzak. 4 irudian klaseen banaketa eta horien arteko erlazioak ikus daitezke.

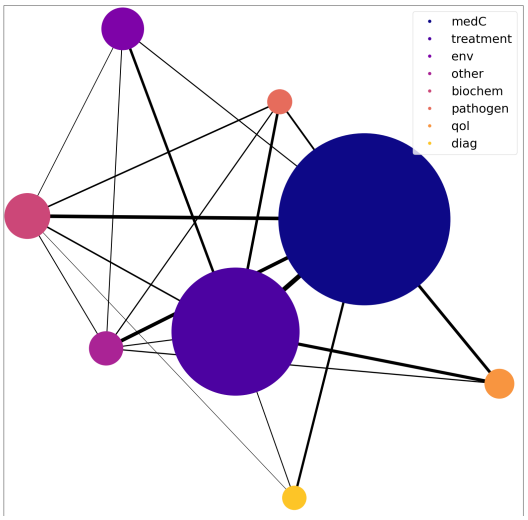

Figure 4: Erabilitako entitateen klaseak tamina adierazita eta arteko erlazioak

5 Emaitzak

Errendimenduaren bilakaera aztertuz gero, bai NERean, bai eredu desberdinen REn, 2 eta 3 tauletan, zeregin horiek sare sozialetan sortutako domeinu biomediko batean egiteak dakarren er-

ronka ikusiko dugu. Arazo nagusia orokortzea da, train ondorengo etapetan huts eginez. Arazo natural bat eskuragarri dagoen datu-kopurua kontuan hartzean, multzoetan gutxitan dauden entitateekin. Erabiltzen duen hiztegi espezifikoa dela eta, ez ziren eraginkorrak izan datuak gehitzeko erabili ziren hainbat teknika.

Erabilitako eredu multzoaren barruan, BioRedditBERTek bi atazetan lan hobea egin zuela erakutsi zuen. BioBERTengandik jasotako ezagutzari eta Reddit bezalako ingurune batean ikasitakoari esker, SciBERTk baino errendimendu handiagoa lortu du, ezagutza eremu zientifikora mugatuta. 5 grafikoko NER nahaste-matrizea ikusita, % 80 inguruko puntuazioak nabaritzen ditugu irudikapen handia duten entitatetan. Bestalde, 6 grafikoko RE nahasteko matrizeak azaltzen du errore nagusiaren iturria harremana osatzen duten entitateetako bat detektatzen ez den kasuei dagokiela.

6 Ondorioak

Lan honetan zenbait hurbilketa egin dira propietate horietako corpus batekin lan egiteak dakarren erronkari aurre egiteko. Bio-medikuntzara eta sare sozialetara bideratutako datu-multzo batekin lan egiteak dituen oztopoak direla eta, klase bakoitzerako lagin gutxi izateaz gain eta normalean erabiltzen diren datuekin alderatuta aniztasun handiagoa izateaz gain, emaitza onak lortu dira. Erronka nagusia eremu horietan lan egitea ahalbidetuko duen hizkuntzaren irudikapena lortzeko beharrezkoak diren hiztegien desberdintasunean datza. Hurbilketa interesgarri bat exBERT ereduan proposatutako soluzioaren egokitzapena litzateke, baliabide gutxiko agertokiekin bateragarria den modulu ikasgarri batekin.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *CoRR*, abs/1909.07755.

Vasiliki Foufi, Tatsawan Timakum, Ma Ba, Christophe Gaudet-Blavignac, Bsc Cs, Christian Lovis, and Min Song. 2019. Mining of textual health information from reddit: Analysis of chronic diseases with extracted entities and their relations. *Journal of Medical Internet Research*, 21:e12876.

John M. Giorgi, Xindi Wang, Nicola Sahar, Won Young Shin, Gary D. Bader, and Bo Wang. 2019. End-to-end named entity recognition and relation extraction using pre-trained language models. *CoRR*, abs/1912.13415.

Yuting Guo, Yao Ge, Mohammed Ali Al-Garadi, and Abeed Sarker. 2021. Pre-trained transformer-based classification and span detection models for social media health applications. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 52–57, Mexico City, Mexico. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammu : A survey of transformer-based biomedical pretrained language models.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Fei Li, Zhichao Lin, Meishan Zhang, and Donghong Ji. 2021. A span-based model for joint overlapped and discontinuous named entity recognition. *CoRR*, abs/2106.14373.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

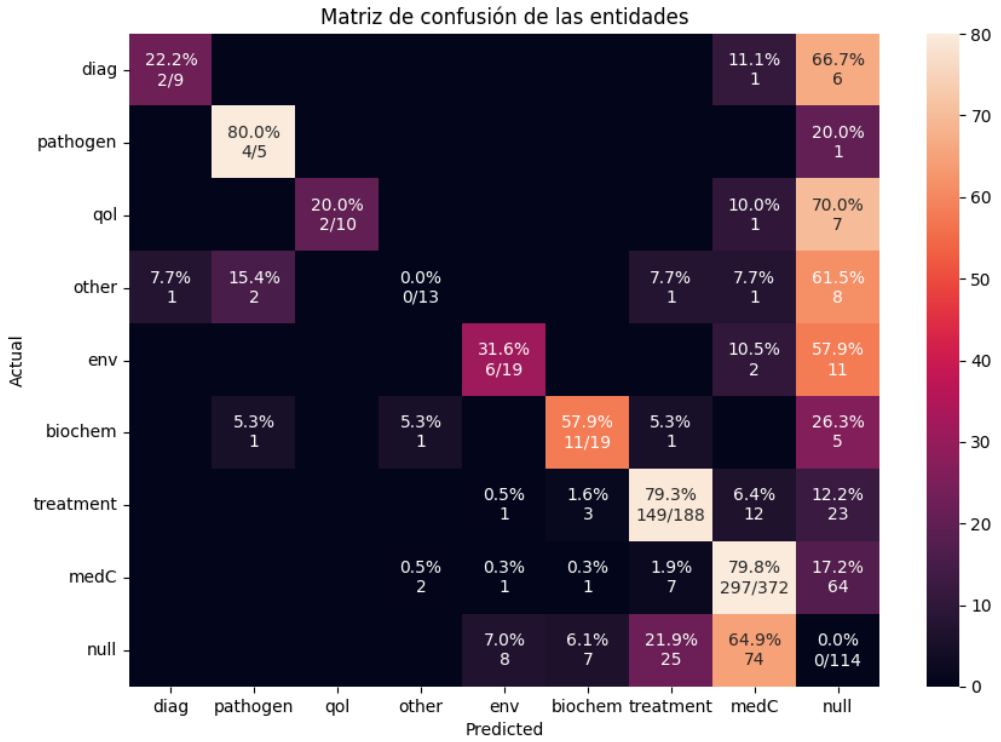

Figure 5: BioRedditBERT NER atazan lortutako konfusio matrizea

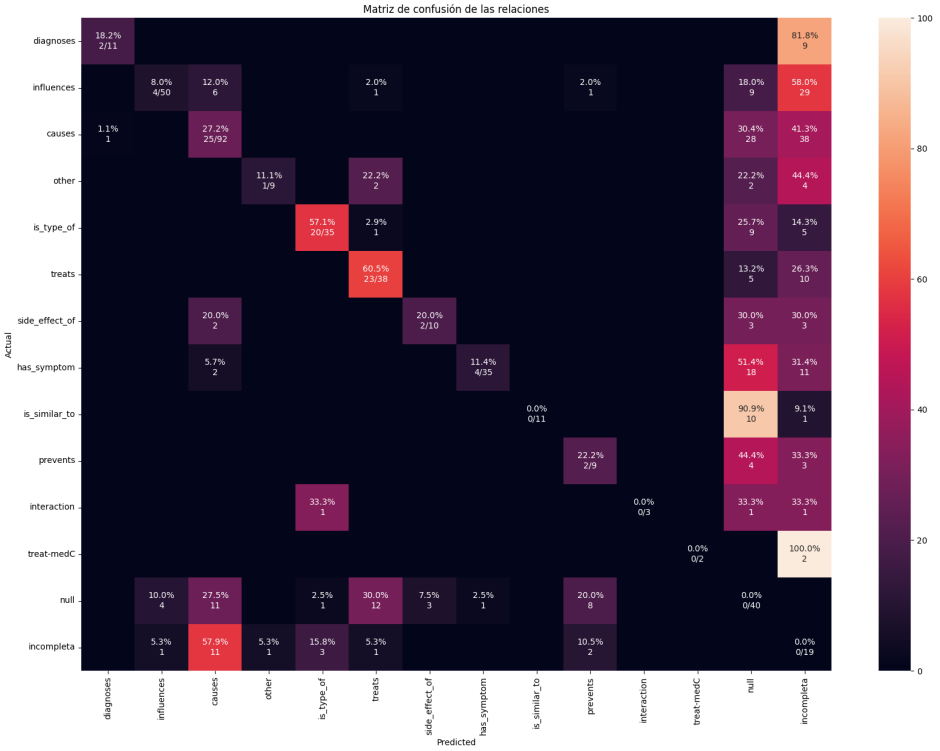

Figure 6: BioRedditBERT RE atazan lortutako konfusio matrizea

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical ner and covid-19 qa.

Sanja Scepanovic, Enrique Martin-Lopez, Daniele Quercia, and Khan Baykaner. 2020. Extracting medical entities from social media. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 170–181, New York, NY, USA. Association for Computing Machinery.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pretrained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2022. Recovering patient journeys: A corpus of biomedical entities and relations on twitter (bear).

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links.

A Supplemental Material

GitHub kodean sartzeko, bidali eskaera helbide honetara: xabilarra96@gmail.com