VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

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

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Suizidio-zantzuen bilaketa eta deskribapena sare sozialetako testuetan

Xabier Irastortza Urbieta,
Maite Oronoz Anchordoqui
eta *Alicia Pérez Ramírez*

93-100 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.11

Suizidio-zantzuen bilaketa eta deskribapena sare sozialetako testuetan

Xabier Irastortza Urbieta, Maite Oronoz Anchordoqui, Alicia Pérez Ramírez

HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU. M. Lardizabal 1. 20080 Donostia.

xabier.irastorza@ehu.eus

Laburpena

Sare sozialak norbere esperientziak adierazteko tresna garrantzitsu bihurtu dira. Esperientzia negatiboak kontatzeko ihesbide ere badira, baita ere gaitz mentalak dituzten pertsonentzat. Lan honek, hain zuzen, sare sozialetan suizidio-zantzuak erakusten dituzten pertsonak ditu jomuga, euren testuetan ezaugarri linguistiko bereizgarriak bilatzen aritu baikara, halako kasuen atzemate automatikoa errazte aldera. Hartarako, euren mezuak oinarritzat hartu eta sistema automatikoen bidez azterketa morfosintaktikoa eta sentimenduen analisia egin dugu. Lexikoa kuantitatiboki alderatu dugu ertz ugaritatik, tartean sistema adimendunak erabiliz. Suizidio-zantzuak adierazten dituzten pertsonek idazterakoan ezaugarri morfosintaktiko bereizgarriak dituztela berresteaz gain, erabiltzaile horiek sentimendu positibo zein negatiboak gai pertsonaletara bideratzeko joera nabarmena dutela ondorioztatu dugu.

Hitz gakoak: Suizidio-ideiagintza, sare sozialak, hizkuntzaren analisia, hizkuntzaren prozesamendua

Abstract

Social networks allow their users to share their concerns and emotions, including those related to mental illness. The aim of this work has been to determine which linguistic features are shown by people with potential suicidal ideation in social networks. For this purpose, we have used a dataset with Twitter messages, and we have analysed them morphosyntactically by means of automatic systems. We have also made several lexical analyses using intelligent systems. We concluded that people with potential suicidal ideation shape their linguistic features in order to express more precisely their inner perceptions and concerns.

Keywords: *Suicidal ideation, social networks, language analysis, Natural Language Procesing*

1 Sarrera eta motibazioa

XXI. mendeko aldaketa sozialen hauspotzaile nabari izaten ari dira sare sozialak. Askotan aipatzen da pertsonen arteko komunikazioa biderkatu dutela edo uneoro nahi adina iturritatik informazioa eskuratzeko ahalmena eman digutela. Alabaina, tresna berritzaile hauei beste hainbat erabilera bilatu zaizkie, adibidez, ez dira pertsona gutxi sare sozialak euren sentimenduak eta hausnarketak azaleratzeko erabiltzen dituztenak, ziur aski, anonimotasunak babestuta edo ingurukoak ez diren pertsonen iritzien bila. Lan honetan erabiltzeko modu horiei heldu diegu, pertsonen barne-munduaren ispilu izan daitezkeelakoan. Horren harira, hainbat ikerlarik (Abd Rahman *et al.*, 2020), erakutsi dute Twitter (egun X deitua), Reddit edo Sina Weibo bezalako sare sozialetan badirela pertsonak euren egoera mentala irekitasunez adierazten dutenak, batzuetan gaixotasun mentalei, jarraitzen dabiltzan tratamenduei edo pairatzen dituzten sintomei erreferentzia eginez. Hortaz, euren aldarteari buruz seriotasunez mintzatzen diren pertsonak baldin badaude sare sozialetan, zergatik ez erabili ingurune hori osasun-erakundeei laguntza emateko? Adibidez, gaixotasun mentalen gorakadak garaiz atzematen identifikatzeko datuak biltzen lagun dezaketela aurreikusten dugu, nahiz eta oraindik bide luzea dagoen egiteko.

Lan honetan suizidio-ideiagintzan —norbere buruaz beste egiteko ideiak izatean— zentratu gara, haren zantzuak erakusten dituztela diruditen pertsonetan, gizartean kezka handia sortzen duten suizidio-kasuen lehen fasea baita. Arlo horretan, beharrezko baimen etikoak edukiz gero, sare sozialetako informazioa funtsezkoa izan daiteke suizidio-kasu indibidualak atzemateko orduan. Baina, noski, sare sozialetan testua soilik dago aztergai eta bertan, kasu gehienetan, pertsona batek ez du esplizituki adierazten suizidio-ideiagintza duela. Horra hor lan honen xede nagusia, suizidio-ideiagintza duen pertsona batek haren idazkeran baliatzen dituen ezaugarri linguistiko bereizgarriak atzeman eta deskribatzea.

2 Arloko egoera eta ikerketaren helburuak

2.1 Aurrekariak

Ez gara ikerketa-lerro honetan murgiltzen garen lehenak, sare sozialen agerpenarekin batera, 2010eko hamarkadatik, ez dira gutxi izan ingurune horietan suizidio-ideiagintza eta baita ere depresioa bezalako gaixotasun mentalak atzemateko lan egin duten ikerlariak. Horietako hainbatek erabiltzaileen metadatuetan jarri izan dute arreta, tartean, mezuen erantzun kopuruak, atsegite kopuruak edo egunero idatzitako mezu kopuruak neurtuz (Abd Rahman *et al.*, 2020). Bide beretik, erabiltzaileen idazte-orduak ere behatu izan dituzte, insomnioa bezalako sintomez oharturik. Hain zuzen, suizidio-ideiagintza duten erabiltzaileek sare sozialak goizaldean gehiago erabiltzen dituztela ondorioztatu zuten De Choudhury *et al.* autoreek (2013) lanean. Ildo beretik, Zogan *et al.* autoreek (2021) lanean erabiltzaileen harreman-sareak aztertu zituzten, besteak beste jarraitzaile kopuruak behatuz eta ondorioztatu zuten depresioa duten erabiltzaileak isolatuago egon ohi direla sare sozialetan.

Alabaina, lan honetan ezaugarri linguistikoetan jarri dugu fokua. Horiek ere aztertu izan dituzte aurrekariek. Ezaugarri linguistikoak topatzeko analizatzaile morfosintaktiko automatikoak erabili izan dira, tartean Spacy[1] edo UDPipe[2], zeintzuek testua tokenetan banatzen duten (hitz ortografikoekin eta puntuazio-markekin bat datozen testu-zatiak) eta token bakoitzeko ezaugarri morfosintaktikoak erauzten dituzten (Gracia Urzelai, 2023). Lexikoa aztertzeko, LIWC[3] tresna ospetsua da ikerketa-lerro honetan, hitzen agerpen-maiztasunetik abiatuta ezaugarri psikolinguistikoak eraugteko gaitasuna duelako (Tausczik eta Pennebaker, 2010). Lexikoiak, ezaugarriren bat partekatzen duten hitzen zerrenda handiak, ere eratu izan dituzte aurrekariek hainbat fenomeno aztertzeko testuetan. Adibidez, Wikipedian oinarrituta, sendagai-izenen lexikoiak sortu izan dira testu-multzoetan lexiko mediko espezializatuaren erabilera neurtzeko (Oyong *et al.*, 2018). Hari beretik, gaixotasunen aipamenak Twitterreko mezuetan automatikoki atzemateko sistemen nazioarteko lehiaketak antolatu izan dira, SocialDisNER, kasurako (Sánchez *et al.*, 2022).

Emozioen analisia eta ezaugarri linguistikoena ezin dira elkarrengandik banandu. Lexikoi espezializatuak erabili izan dira emozioen azterketa zehatza egiteko, horiek osatzen dituzten hitzak emozio zehatzen adierazpenarekin eta beste faktore psikologikoekin hertsiki lotzen direlako. Arau afektiboei lotutako ANEW lexikoia ezaguna da (Bradley eta Lang, 1999). Hala ere, azterketa maila sinpleagoan egiteko, sentimendu-balentziaren edo oinarrizko sei emozioen mailan, ikaskuntza sakoneko ereduak dira erabilienak egun, batez ere Transformer arkitekturadun (Vaswani *et al.*, 2017) eredu espezializatuak (Pérez *et al.*, 2021). Azken urteetan Hizkuntza-eredu Handiak ere erabiltzen hasiak dira ataza horietarako (Hanafi *et al.*, 2024).

Aurrekariek ondorioztatu izan dute suizidio-ideiagintza duten pertsonek lehen pertsonako izenordainak maizago eta hirugarrenekoak urriago erabiltzen dituztela (De Choudhury *et al.*, 2013). Gainera, sendagaien izen gehiago aipatzen dituztela (Oyong *et al.*, 2018) eta sentimendu negatibo indartsuagoak erakusten dituztela ere aipatu izan dute (De Choudhury *et al.*, 2013). Euskal Herriko Unibertsitatean ere egin izan dira zenbait lan ikerketa-lerro honetan. Gracia Urzelai autoreak (2023) lanean lehen pertsonako izenordainen maiztasun handiagoa topatu zuen Reddit sare-sozialeko suizidio-zantzudun ingelesezko testuetan eta sailkatzaile automatikoak probatu zituen, Transformer arkitekturan eta LDA ereduetan oinarritutakoak, halako testuak automatikoki identifikatzea helburu izanda. Ostera, Oronoz *et al.* autoreak (2024) lanean gaztelaniazko testuak aztertuz ere ondorio berera heldu ziren.

2.2 Ikerketaren helburuak

Ikerketa honetan suizidio-ideiagintzak testuetan zein adababki gehitzen dituen, zein lorratz uzten dituen zehaztasunez deskribatzen saiatu gara, datu-bilduma handi bat eratu ondoren, bertako testuak sistema automatikoen bidez hizkuntzaren bi alderditan aztertuz:

- Maila morfosintaktikoa. Suizidio-zantzuak erakusten dituzten pertsonen testuetan zein ezaugarri morfosintaktiko bereizgarri ageri diren aurkitzen eta haien bereizgarritasuna kuantifikatzen saiatu gara.
- Maila lexikoa. Suizidio-zantzuak erakusten dituzten pertsonen ezaugarri lexiko bereizgarriak topatzen ere saiatu gara, hainbat eremu semantiko aztertuz, tartean, gaixotasun-izenak, sendagai-izenak eta lanbide-izenak. Mezuaren emozionaltasunaren eta kategoria gramatikalaren arabera ere egin dugu azterketa lexikoa.

[1] spaCy aztertzaile morfosintaktikoa: `https://spacy.io/`
[2] UDPipe aztertzaile morfosintaktikoa: `https://lindat.mff.cuni.cz/services/udpipe/`
[3] Ingelesezko sigletatik, *Linguistic Inquiry and Word Count*: `https://www.liwc.app/`

Azterketa horiek egiteko suizidio-zantzuak erakusten dituzten pertsonen testuak eta halakorik erakusten ez dituztenak alderatu ditugu etengabe. Gure datu-bilduma gaztelaniazko testuek osatzen dute eta ez, ikerketa-lerroan ohikoa den bezala, ingelesezkoek. Hori da lan honen ekarpenetako bat eta horrek ekarpen nagusia izan litekeena are gehiago handitzen du: sistema automatikoen sorkuntza alboratu eta ezaugarri linguistikoen analisi sistematikoan ardaztu den lana izatea, ikerketa-lerroan ohikoa denaren alderantzira.

3 Ikerketaren muina

Atal honetan, lehenengo, datuak nola eskuratu ditugun azalduko dugu eta ostera, egindako bi analisien (morfosintaktikoa eta lexikoa) metodologia azaldu eta emaitzak eztabaidatuko ditugu.

3.1 Datuak

2023an Twitter sare sozialean suizidio-zantzuak erakusten dituzten erabiltzaileen mezu-bilduma bat sortzeko lanak egin genituen. Hartarako, 98 esaldi-gako eratu genituen (adibidez, euskarara itzulita, *nire bizitzak ez du zentzurik* edo *bizitza gorroto dut*), suizidio-ideiagintza duen pertsona batek noizbait esan litzakeen esaldi edo hitzak irudikatzen dituztenak. Twitter-ek eskainitako Aplikazioak Programatzeko Interfazea (API) baliatuta, azken hilabetean gutxienez esaldi-gako horietako bat idatzi zuten erabiltzaileak bilatu genituen sare sozialean barrena. Hala, 6.451 mezu topatu genituen eta ondoren, mezu horiek eskuz aztertzeari ekin genion, zalantzazko kasuetan erabiltzaileen profilak ere begiratzeraino, jakiteko ea mezu bakoitza idatzi zuen erabiltzaileak itxuraz suizidio-zantzuak erakusten ote zituen (edo, aitzitik, beste testuinguru batean idatzitako mezuak ziren). Erabaki horren arabera, bi erabiltzaile-multzo sortu genituen: zantzuak erakusten zituzten erabiltzaileena (positiboak, 957 erabiltzaile) eta ez zituztenena (negatiboak, 2.950 erabiltzaile), alderaketak egin ahal izateko. Azkenerako, mezu horien egileen profiletatik ahalik eta mezu gehien jaitsi genituen eta horrela eratu zen lan honetan oinarritzat hartu dugun datu-bilduma, ia milioi bat mezuz osatua. Jasotako datu-sortaren ezaugarriak bildu ditugu 1. Taulan.

1. Taula: Datu-bildumaren dimentsioak. Hurrenez hurren, multzo bakoitzeko erabiltzaile kopurua, erabiltzaile bakoitzeko batez besteko mezu kopurua, mezuko batez besteko token kopurua eta multzo bakoitzeko guztizko mezu kopurua eta token kopurua. Normalean, token bakoitzak hitz ortografiko bat adierazten du.

Erabiltzaile multzoa	Erabiltzaile kopurua	Mezuak erabiltzaileko	Tokenak mezuko	Mezu kopurua	Token kopurua
Positiboak	957	169,53	10,29	162.238	1.669.764
Negatiboak	2.950	280,17	10,71	826.504	8.848.743
Orotara	3.907	253,07	10,64	988.742	10.518.507

Erabili dugun datu-bilduma desorekatuta dago bai erabiltzaile kopuruarekiko eta baita ere mezuko kopuruarekiko, handiagoa da bi kasuetan erabiltzaile-multzo negatiboa (ikus 1. Taula). Horregatik, lanean zehar emaitzak termino erlatiboetan aurkeztu dira. Lanean zehar egin diren analisi guztietan bi multzo horiek alderatu dira.

3.2 Morfosintaxiaren azterketa

Suizidio-ideiagintzak pertsona batek idatzitako testuen morfosintaxian uzten dituen lorratzak ikertu ditugu atal honetan, lehenik baliatutako metodologia azalduz eta jarraian lortutako emaitzak eztabaidatuz.

3.2.1 Metodologia

Aztarna morfosintaktikoak ikertzeko datu-bilduma osoan aztertzaile morfosintaktiko bat aplikatu dugu, testuak tokenetan banatu eta token bakoitzari dagozkion ezaugarri morfosintaktikoak erauzteko (hala nola, generoa, pertsona, denbora edo funtzio sintaktikoa). Spacy-UDPipe tresna erabili dugu, zeinak testua tokenizatzeko eta ondoren token bakoitzari analisi morfosintaktikoa egiteko prozesaketa-kate bat eskaintzen duen. Tresna horrek 227 ezaugarri morfosintaktiko ezberdin aztertzeko aukera ematen du token bakoitzeko eta ezaugarri horien guztien balioak kalkulatu dira lan honetan. Aurrekariek (Oronoz *et al.*, 2024) baliatu izanak eraman gaitu tresna aukeratzera.

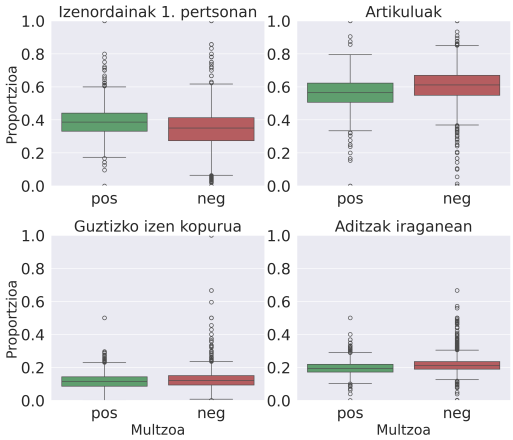
Behin tokenen ezaugarriak jasota, horiek kuantifikatzea izan da ondorengo lana. Lehenik, testuak euren egileen arabera sailkatu ditugu, erabiltzaile bakoitzeko testu-multzo bat edukitzeko. Ondoren, testu-multzo horietan ezaugarri bakoitzaren agerpenak zenbatu ditugu eta noski, testu-multzo batzuk besteak baino handiagoak direnez,

1. Irudia: Ezkerreko taulan ezaugarri bakoitzaren agerpen-proportzioa ikusgai, erabiltzaile-multzoaren arabera. Hartarako, erabiltzaile bakoitzaren testuetako agerpen-proportzioa kalkulatu da lehenik, horien batez bestekoa egiteko gero (ikus 3.2.1 Atala). Eskuinean, erabiltzaile mailako agerpen-proportzioen banaketa, lau ezaugarritan.

(a) Ezaugarri morfosintaktiko bereizgarriak.

Ezaugarriak	Bat. bez. erab. pos.	Bat. bez. erab. neg.
Izenordain singularrak	0,73	0,67
Izenordainak 1 pertsonan	0,39	0,33
Laguntzaileak 3. pertsonan	0,50	0,55
Mugatzaile artikuluak	0,56	0,60
Aditzak 3. pertsonan	0,39	0,43
Aditzak pluralean	0,10	0,13
Izen kopurua	0,20	0,21
Aditz kopurua	0,15	0,14
Laguntzaileak orainaldian	0,67	0,68
Aditzak iraganean	0,12	0,12
Laguntzaileak iraganean	0,05	0,05

(b) Ezaugarri morfosintaktikoen bereizgarritasuna.



alderaketa justuak egiteko, zenbaketak erlatibizatu ditugu. Hau da, erabiltzaile bakoitzak ezaugarri bakoitza zein maiztasun erlatiborekin baliatzen duen kalkulatu dugu. Jarraian, erabiltzaile bakoitzaren multzoa (positiboa edo negatiboa) aintzat hartuta, maiztasun erlatibo horien batez bestekoak kalkulatu ditugu, erabiltzaile-multzoen ordezkari. Azken maila honetan, erabiltzaile-multzo bakoitzeko 227 balio lortu ditugu, ezaugarri bakoitzeko bana eta balio horiek erabili ditugu multzoak elkarren artean alderatzeko.

3.2.2 Emaitzak eta eztabaida

Emaitzak aurkezteko, ezaugarri morfosintaktikoen inguruan izan ditzakegun zenbait aurreiritzi eta usteetan ardaztuko dugu atal hau. Hasteko, zaila egin dakioke irakurleari sinistea pertsona baten hizkuntzaren egitura aldatu egin daitekeela egoera mentalak eraginda, kasu honetan suizidio-ideiagintzak. Ateratako emaitzek ikuspuntu horri indarra kendu eta eman egiten diote, aldi berean. Kendu, badirelako zenbait ezaugarri erabiltzaile positiboetan ugariagoak direnak erabiltzaile negatiboekin alderatuta (ikusi 1a Taula). Indarra eman, erabiltzaile-multzoen arteko desberdintasunak txikiak direlako eta estatistikoki esanguratsuak direla ezin delako esan (ikusi 1b Irudia). Hala ere, kontuan hartu behar da desberdintasun horiek estatistikoki esanguratsuak izan ez arren, koherentzia izan badutela eta aurrekariek beste datu-bildumetan aurkitutakoekin bat datozela (De Choudhury *et al.*, 2013; Gracia Urzelai, 2023). Ezberdintasun txiki horiek joera argi bat erakusten dute, esan bezala, 1a Taulan islatua. Taula hori errazago ulertzeko adibide bi: lehen errenkadak erakusten du erabiltzaile positiboek izenordainen % 73 singularrean idazten dituztela batez beste, negatiboek aldiz % 67; aditzen kasuan, positiboetan batez beste token guztien % 15 aditzak direla adierazten du taulak.

Ildo horri jarraikiz, logikoa egingo litzaiguke pentsatzea suizidio-ideiagintza duten pertsonak zentratuago egongo direla beraien barne munduan, beraien kezketan, ideiagintza horren sorburu baitira. Beraz, ez litzateke arraroa izango lehen pertsona gramatikala eta numero singularraren erabilera altuagoa izatea (nitasunaren presentzia) eta aldiz, norbera kide den taldeekiko erreferentzia gutxiago egitea (hirugarren pertsona singularraren erabilera baxuagoa) eta ziur aski, kanpoko taldeei ere aipamen gutxiago egitea (hirugarren pertsona plurala). Emaitzek argi erakusten dute hori (1a Taula).

Bestalde, pentsa dezakegu suizidio-ideiagintza duten pertsonek ezaugarri morfosintaktiko bereizgarriak izanda, denbora gramatikalak zeresana izango duela horietan. Zehazki, iraganeko aditzak ugariagoak izango direla etorkizunekoak baino, pertsona hauek haien iraganeko arazoetan zentratuago daudela esan ohi delako. Haatik, lortu ditugun emaitzek ez dute ebidentzia sendorik azaleratu zentzu horretan, desberdintasunak baztergarriak dira erabiltzaile-multzoen artean. Baliteke testu-motak (sare sozialetako mezu espontaneoak eta horien berehalakotasunak) eragina izatea edo, agian, ezaugarri honek hasiera batean pentsa daitekeen pisu handi hori ez izatea.

Bada beste hipotesi bat logikoa iruditu dakigukeena: suizidio-ideiagintza duten pertsonek ekintza gutxiago deskribatuko dituztela haien testuetan, egunerokotasunean aktibitate txikiagoa izatearen edo isolamendu sozialaren ondorioz. Hipotesia zuzena balitz, erabiltzaile positibo eta negatiboek erabilitako aditz kopuruan desberdintasunak

iguriki beharko genituzke. Kasu honetan ere, emaitzek ezin izan dute hipotesi hori baieztatu, desberdintasunak baztergarriak direlako berriz ere. Aditz eta izen kopuru beretsua dute bi multzoetako erabiltzaileek. Azalpena aditz zein izenen aniztasun handian egon daiteke eta baliteke aniztasun horretan barneratu ahala desberdintasunak agertzea, adibidez, aditz abstraktu eta konkretu kopurua neurtuz gero. Bide horretatik sakondu dugu 3.3 atalean.

3.3 Lexikoaren azterketa

Suizidio-ideiagintza duten pertsonek lexiko bereizgarririk erabiltzen ote duten ere galdegin diogu gure buruari. Galdera horri erantzuteko hiru bide jorratu ditugu: lexiko espezializatuaren azterketa, sentimenduen araberako lexikoaren azterketa eta kategoria gramatikalen araberako lexikoaren azterketa.

3.3.1 Metodologia

Lexiko espezializatu gisa ulertu dugu psikologiarekin edo medikuntzarekin lotutako termino-sorta, adibidez, sendagaien izenak, arlo horiekin lotutako lanbideenak, gaixotasunen izenak edo sintomenak. Bi bide jarraitu ditugu termino horien erabilerari buruzko informazioa lortzeko: lexikoi baten eraketa eta sistema adimendun baten erabilera. Lexikoiaren bidez gaixotasunen izenen eta gaixotasunekin lotutako termino medikoen (sintomak, kausak, sendagaien izenak eta ondorioak) agerpen kopuruak kalkulatu ditugu. Hartarako Wikidata[4] erabili dugu, bertan termino horiek grafo moduan egituratuta egonik, errazagoa baita termino egokiak nahi den hizkuntzan eskuratzea. Datu-bilduma hori baliatuta 369 terminoz osatutako lexikoi bat eratu dugu, psikologia edo psikiatriarekin lotura duten gaixotasun-izenak ardatz hartuta eta grafoa baliatuz, horien sintoma, sendagai eta bestelako erlazionatutako terminoak ere eskuratuz. Termino horien adibide, besteak beste: *alprazolam*, *agomelatina*, *antsietatea*, *zefalea* eta *alopezia*. bezalako sintoma-izenak daude lexikoi horretan. Lexikoi horretako hitz bakoitza erabiltzaile-multzo bakoitzean zenbat aldiz ageri den kalkulatuz egin ditugu alderaketak.

Aldiz, lanbide-izenen agerpenak atzemateko beste hurbilpen bat erabili dugu. Carrasco eta Rosillo autoreek (2021) lanean ProfNER lehiaketarako sortutako sistema adimenduna erabiltzea erabaki dugu, zeinak, hitzbektoreak eta kosinu-antzekotasun prozedura bat konbinatzen dituen, eta era horretan, testu bat emanda automatikoki identifikatzen dituen lanbide-izenak, gu bila gabiltzan espezializatuak izan (psikiatria, psikologia edo medikuntza arlokoak) edo beste edozein. Tresnaren emaitza txukunek eta dokumentazio abegikorrak eraman gaitu hura erabiltzera. Modu honetan, aukera izan dugu edozein arlotako lanbide-izenekin alderaketak egiteko erabiltzaile positibo eta negatiboen artean. Guztira, 15.711 lanbide-izen erauztea lortu dugu datu-bilduma osoan. Kasu honetan ere lanbide-izenen agerpen kopuruak alderatu ditugu erabiltzaile-multzoen artean.

Lexikoaren erabilerak sentimenduekin zerikusia izan dezakeela suposatu dugunez, datu-bildumako mezu bakoitzean sentimenduen analisia egitea erabaki dugu. Hartarako, Pérez *et al.* autoreek (2021) lanean aurkeztutako *Pysentimiento* izeneko tresna baliatu dugu, zeina testu bat emanik bertan oinarrizko sei emozioetatik (alaitasuna, goibeltasuna, haserrea, beldurra, higuina eta harridura) zein nagusitzen den inferitzeko gai den. Hori lortzeko, BERT Transformer arkitekturan (Devlin, 2018) oinarritutako sistema adimendun bat baliatzen du, autoreek gaztelaniazko eta portugesezko testuekin entrenatuta. Hain zuzen, tresna gaztelaniazko testuekin entrenatuta egoteak eta haren eskuragarritasunak eraman gaitu hura baliatzera. Gailu honi esker mezuak haietan nagusitzen den emozioaren arabera sailkatzeko gai izan gara eta horrela, mezu-multzo horien artean lexikoa alderatzeko atea ireki zaigu. Esandakoagatik, analisi honetan 6 mezu-multzo sortu zaizkigu erabiltzaile-multzo bakoitzeko. Lexikoa alderatzerakoan, lema bakoitzaren agerpen-kopurua zenbatu dugu mezu-multzo bakoitzean eta hitz-hutsak (hizkuntza batean ohikoenak direnak) bahetu egin ditugu hasiera-hasieratik.

Azkenik, kategoria gramatikalaren arabera lexiko erabiliena zein den ere jakin nahi izan dugu. 3.2.2 atalean planteatutako zalantzak motibatuta, soilik izen eta aditzak hartu ditugu aintzat. Hartarako, 3.2.1 atalean egindako analisi morfosintaktikoa berrerabili dugu. Lemak izen edo aditz gisa sailkatu eta bakoitzaren agerpen-kopuruak zenbatzeari ekin diogu. Era horretan, erabiltzaile-multzo bakoitzeko aditz eta izen forma guztien zenbaketak lortu ditugu, horiekin alderaketak egiteko. Kasu guztietan, erabiltzaile-multzoen artean agerpen-kopuru desberdinena duten forma lexikoak aurkitzeko, lehenik baheketa bat egin dugu, aztertu nahi izan dugun baldintza (adibidez, tristura-mezuak izatea edo soilik aditzak kontuan hartzea) baten bidez, eta ondoren, aukeratutako testuak bi testu-multzo kontrajarritan metatu ditugu (erabiltzaile positiboenak eta negatiboak). Testu-multzo horien artean egin dugu alderaketa, agerpen-kopuru ezberdinenak dituzten forma lexikoak bilatuz, beti ere termino erlatiboetan arituta (ikus 1. Taulako datu-desoreka).

[4]Bertako SPARQL kontsulta-zerbitzua, hain zuzen: `https://query.wikidata.org/`

3.3.2 Emaitzak eta eztabaida

Morfosintaxiarekin alderatuta, irakurleak pentsa dezake desberdintasun lexikoak ohikoagoak izan daitezkeela suizidio-ideiagintza duten pertsonen idazkeran, azken finean, euren sentimenduak adierazten badituzte, hitz jakin batzuen erabilera ezinbestekoa izango delako. Atal honetan uste hori datuen bidez baieztatzen saiatzeaz gain, ezaugarri lexiko horiek zehatzago deskribatzen saiatu gara.

2. Irudia: Aditz eta izenen analisia. Ezkerrean, aditz eta izen ugarienak erabiltzaile-multzo bakoitzean, hitz-hutsak kenduta. Eskuinean, aditz eta izen nabarienak, 3.3.1 atalean azaldutako prozedura jarraituz.

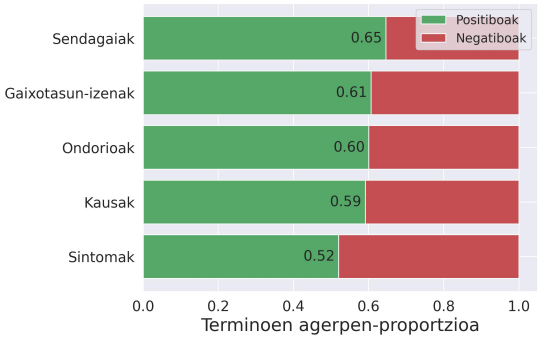
(a) Aditz eta izen ohikoenak

(b) Aditz eta izen nabarmenduak

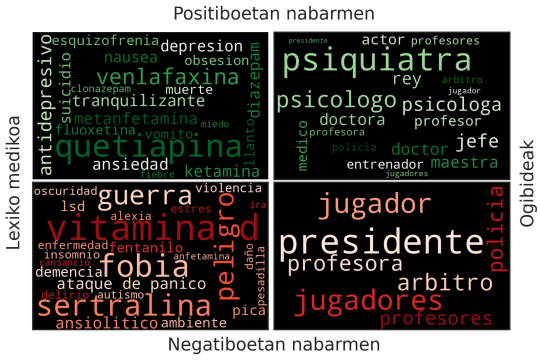
Hasteko, 3.2.2 atalean planteatu dugun zalantzari helduko diogu berriz. Atal horretan ikusi dugu erabiltzaile-multzoen artean ez dagoela desberdintasun handirik aditz eta izenen maiztasunean, biak ere antzera erabiltzen dituzte erabiltzaile guztiek. Arestian planteatu dugu, ordea, aditz edo izen zehatzen maiztasunak aztertuz gero, agian, ezberdintasunak azaleratu litezkeela. Bide horretan, 2a Irudiak argi erakusten du aditz eta izen ohikoenak berdinak direla bi erabiltzaile-multzoetan, hala nola, *querer*, *tener* eta *gente*. Alabaina, erabiltzaile-multzoen artean maiztasun desberdinenak erakusten dituzten hitzak azaleratuz, irudi erabat desberdina lortu dugu (2b Irudia). Irudi horretan, barne egoerari eta ekintza pertsonalei erreferentzia egiteko izen eta aditzak dira nagusi erabiltzaile positiboetan (tartean *valgo* edo *ayuno*) eta aldiz, gai sozialagoei eta kanpo munduari lotutakoak erabiltzaile negatiboetan (adibide gisa *votar* edo *presidente*).

3. Irudia: Lexiko espezializatuaren analisia. Ezkerrean, medikuntza-termino mota bakoitzaren agerpen-kopuruaren alderaketa bi erabiltzaile-multzoen artean, termino erlatiboetan. Termino-mota bakoitzaren agerpenen zein proportzio diren erabiltze positiboenak adierazten dute zenbakiek, multzo bakoitzeko testu-kopuruaren arabera normalizazioa egin ostean. Eskuinean, termino mediko eta lanbide-izen nabarmenenak erabiltzaile-multzo bakoitzean (ikus 3.3.1 atala).

(a) Medikuntza-terminoen agerpena, erlatiboki



(b) Hitz nabarmenenak erabiltzaile-multzo bakoitzean

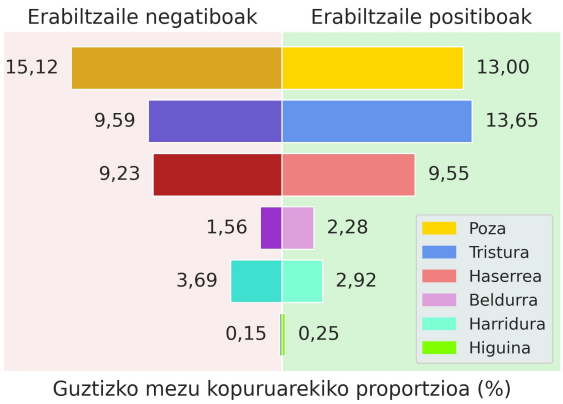


Bestalde, aurrekariek aipatu izan dute medikuntza edo psikologiarekin lotutako hitz gehiago erabiltzen dituztela suizidio-zantzuak erakusten dituzten pertsonek (Oyong *et al.*, 2018). Aztergai dugun datu-bilduman 3a Irudian ageri diren emaitzak atera ditugu eta ondorioztatu erabiltzaile positiboek lexiko medikoa ugariago erabiltzen dutela

(bereziki sendagai-izenak), hala ere, erabiltzaile negatiboek ere erabiltzen dituzte eta ez eskaski. Baliteke, ordea, erabiltzaile-multzoen arteko desberdintasun esanguratsuena erabiltzen dituzten termino zehatzetan egotea, 3b Irudian ikusten den moduan. Bertan, erabiltzaile-negatiboetan egunerokotasunean ohikoagoak diren terminoak ageri dira nabarmenduta (*vitamina d*, *fobia*, *peligro*, etab.) eta positiboetan termino zehatzagoak dira gailentzen direnak (*quetiapina*, *venlafaxina* eta sendagai-izenak bereziki). Lanbide-izenetan ere joera bera ikusten da: psikologia eta medikuntzarekin lotutakoak dira nabarmenenak erabiltzaile positiboetan eta lanbide orokorrak negatiboetan.

4. Irudia: Emozioen analisiaren laburpena. Ezkerrean erabiltzaile-multzo bakoitzean emozio bakoitza mezuen zein ehunekotan den nagusi. Mezu gehienei ezin izan zaie inferitu gailentzen den emozioa. Eskuinean, alaitasuna edo tristura gailendu diren mezuetan nabarmentzen diren hitzen arteko alderaketa, erabiltzaile-multzoen artean.

(a) Emozioen agerpena ehunekotan



(b) Lema nabarmenak emozioen arabera

Emozioak adierazteko erabiltzen den lexikoa aztertzeak gehiago argitu dezake esku-artean dugun gaia. Argiemaile dugu 4a Irudia. Bertan, erabiltzaile positibo eta negatiboek emozioak haien mezuetan antzeko maiztasunez adierazten dituztela ikusi daiteke, nahiz eta positiboek tristura zertxobait maizago (% 4,06 gehiago) eta alaitasuna apur bat gutxiago (% 2,12 gutxiago) adierazi. Edonola ere, esan behar da mezu gehienetan sistema ez dela gai izan emoziorik eraunzteko (hurrenez hurren, mezuen % 58,4 eta % 60,7tan). Desberdintasun nabarmenenak, baina, emozioak adierazteko erabilitako lexikoan aurki daitezke, 4b Irudia lekuko. Bertan beha daiteke alaitasun eta haserre mezuetan erabiltzaile-multzo bakoitza nabarmentzen duten hitzak zeharo desberdinak direla. Patroi orokor gisa, esan daiteke positiboek idazterakoan emozio horiek beraien barne mundura edo arazo pertsonaletara bideratzeko joera dutela (besteak beste, *novio* edo *inseguridad* hitzek erakusten duten bezala) eta kontrakoa negatiboek, gizarte mailako gaietara bideratzen baitute euren adierazpena (horren erakusle *selección* edo *empresa* hitzak).

4 Ondorioak

Lan honek suizidio-zantzuak erakusten dituzten erabiltzaileen testuek ezaugarri linguistiko bereizgarriak dituztela berretsi du, aurrekariek hasitako bidea jarraituz. Gure gaztelaniazko testuetan ingelesean aurkitu diren hainbat ezaugarri topatu dira: lehen pertsonako izenordainen maiztasun altuagoa, numero singularraren maiztasun handiagoa hainbat kategoria gramatikaletan, lexiko medikoaren erabilpen handiagoa, etab. Baina urrunago iristea ere lortu du, suizidio-zantzuak erakusten dituzten erabiltzaileek sare sozialetan parte-hartzeko era desberdina dutela azpimarratuz. Modu bereizgarri hori hainbat alderditan hezurmamitzen dela ondorioztatu da: egunerokotasuneko lexikotik urruntzen diren hitzen ugaritasunean (psikiatriarekin lotutako sendagai eta gaixotasun izen gehiago), psikologiarekin lotutako lanbide-izenen ugaritasunean, emozioak bideratzeko erabilitako lexiko desberdinean edo eremu semantiko jakinekin lotutako aditz eta izenen ugaritasunean. Funtsean, suizidio-zantzuak erakusten dituzten erabiltzaileek euren barne mundua edo esperientzia pertsonalak adierazteko baliabide linguistikoak joriago erabiltzen dituztela ematen du eta aitzitik, halako zantzurik erakusten ez duten erabiltzaileetan gizarte-gaiak edo esperientzia pertsonaletik urruntzen direnak adierazteko baliabide linguistikoak nabarmenagoak direla dirudi.

5 Etorkizunerako planteatzen den norabidea

Euskal Herriko Unibertsitateko ikerlariek lehenengo ingelesezko testuak aztertu zituzten, ondoren gaztelaniazkoekin hasi eta lan honen bidez horretan jarraitu dugu, baina bide-orri hau burutzeko euskarazko testuen analisia dago

egiteke. Ikerketa-lerro honetan ingelesezko testuekin egin da lan gehienbat, ikerkuntza munduan hizkuntza horrek duen pisuak eraginda bereziki, baina baita ere hizkuntza horretan dagoen datu-ugaritasunak bultzatuta. Gaztelaniaz datu-faltarik nabaritu ez dugun arren, euskarazko testuetan datu-falta hori agerikoa izatea aurreikusten dugu eta ziur aski, hori izango da bide honi jarraipena emateko oztopo nagusia. Alabaina, hiru hizkuntza horietan analisia egiteko aukera balego, hizkuntza batetik bestera lan honetan zehar deskribatutako ezaugarriak aldatzen ote diren eta baiezkoan, nola aldatzen diren ikertzeko aukera legoke. Gainera, hiru hizkuntza horiek jatorri ezberdinekoak izanda, euskararen kasuan zeharo ezberdina, hizkuntza-arteko analisi hori bereziki erakargarria litzateke maila linguistikoan. Halaber, euskal hiztun elebidunek lan honetan aztertutako esperientziak nola adierazten dituzten ikertzea interesgarria litzateke, diglosiaren eragina agerrarazte aldera.

Erreferentziak

Abd Rahman, Rohizah, Khairuddin Omar, Shahrul Azman Mohd Noah, Mohd Shahrul Nizam Mohd Danuri, eta Mohammed Ali Al-Garadi. 2020. Application of machine learning methods in mental health detection: a systematic review. *IEEE Access* 8.183952–183964.

Bradley, Margaret M, eta Peter J Lang. 1999. Affective norms for english words (anew): Technical manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida* .

Carrasco, Sergio Santamaría, eta Roberto Cuervo Rosillo. 2021. Word embeddings, cosine similarity and deep learning for identification of professions & occupations in health-related social media. In *Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, 74–76.

De Choudhury, Munmun, Michael Gamon, Scott Counts, eta Eric Horvitz. 2013. Predicting depression via social media. In *International AAAI conference on web and social media*, volume 7, 128–137.

Devlin, Jacob. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Gracia Urzelai, Sara. 2023. Suizidio-ideiagintzaren identifikazioa sare sozialetan. *ADDI:* `https://addi.ehu.es/handle/10810/63221` .

Hanafi, Abdelrahman, Mohammed Saad, Noureldin Zahran, Radwa J Hanafy, eta Mohammed E Fouda. 2024. A comprehensive evaluation of large language models on mental illnesses. *arXiv preprint arXiv:2409.15687* .

Oronoz, Maite, Sara Gracia, Jose Mari González, eta Alicia Pérez. 2024. Suizidio-zantzuak sare sozialetan: hizkuntza-ezaugarriak berdinak al dira ingelesez eta gaztelaniaz? *Ekaia* Hurrengo argitalpenak 2024.

Oyong, Irwan, Ema Utami, eta Emha Taufiq Luthfi. 2018. Natural language processing and lexical approach for depression symptoms screening of indonesian twitter user. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 359–364. IEEE.

Pérez, Juan Manuel, Mariela Rajngewerc, Juan Carlos Giudici, Damián A Furman, Franco Luque, Laura Alonso Alemany, eta María Vanina Martínez. 2021. pysentimiento: a python toolkit for opinion mining and social NLP tasks. *arXiv preprint arXiv:2106.09462* .

Sánchez, Luis Gasco, Darryl Estrada Zavala, Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, eta Martin Krallinger. 2022. The socialdisner shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora. In *The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, 182–189.

Tausczik, Yla R, eta James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29.24–54.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanÑ. Gomez, Łukasz Kaiser, eta Illia Polosukhin. 2017. Attention is all you need. In *31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Zogan, Hamad, Imran Razzak, Shoaib Jameel, eta Guandong Xu. 2021. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878* .

6 Eskerrak eta oharrak

Lan hau Eusko Jaurlaritzak (IXA IT-1570-22) eta LOTU ikerketa-proiektuaren (TED2021-130398B-C22) bidez MCIN/AEI Zientzia eta Berrikuntza Ministeritzak eta Europar Batasunak diruz lagundu dute.