EKAIA (202x), artikulua prentsan/article in press
https://doi.org/10.1387/ekaia.xxxxxx
Behin-behineko bertsioa (euskara-orrazketaren faltan).

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN: 0214-9001
e-ISSN: 2444-3255
©UPV/EHU Press

Suizidio-zantzuak sare sozialetan: ingelesez eta gaztelaniaz hizkuntza-ezaugarriak berdinak al dira?

(Traits of suicide in social media: are linguistic cues similar in English and in Spanish?)

Maite Oronoz, Sara Gracia, Jose Mari González, Alicia Pérez*
HiTZ Hizkuntza Teknologiako Euskal Zentroa - Ixa Taldea, Euskal Herriko Unibertsitatea (UPV/EHU)

LABURPENA: Lan hau adimen artifizialeko hizkuntzaren prozesamendua izeneko alorrean kokatzen da eta helburua du, sare sozialetako testuetan agertzen diren eta suizidioarekin zerikusia duten ezaugarri psiko-linguistikoak aztertzea. Sistema adimentsuek suizidio-ideiagintzarekin lotutako ezaugarriok automatikoki erauzteko gaitasuna dute eta kasu honetan, aurrekariek aipatutako ezaugarri horietako batzuk (mezuen luzera, aditzen denborak etab.) aztertu ditugu bi sare sozialetan eta bi hizkuntzetan: ingelesez Reddit-en eta gaztelaniaz Twitter-en. Ezaugarri psiko-linguistikoak *suizidioa* eta *ez-suizidioa* klasetako mezuetan alderatu dira, baita hizkuntzen artean ere. Emaitzei dagokionez, suizidioarekin zerikusia duten mezuetan lehen pertsona singularreko izenordain gehiago erabiltzen dela egiaztatu da bi hizkuntzetan. Ingelesez idatzitako datu-sortan ondorioztatu da suizidio ideiagintzaren zantzuak erakusten dituzten mezuak luzeagoak direla baina ez da hala gertatu gaztelaniazko bilduman, segur aski Twitterren luzera muga dela-eta. Galdera-marka kopurua eta aditzen denboren erabilera, aurrekarietan ez bezala, ez dira esanguratsuak bi klaseak bereizteko.
HITZ GAKOAK: Hizkuntzaren prozesamendua, suizidioa, sare sozialak, hizkuntza-ezaugarriak

ABSTRACT: *This research work is in the field of artificial intelligence called language processing and our goal is to analyse psycho-linguistic characteristics related to suicide in social networks. Intelligent systems have the ability to automatically extract this kind of characteristics associated with suicide ideation. In this work we investigate some suicide related characteristics mentioned in the antecedents (length of messages, verb tenses etc.) in two social networks in two languages: in Reddit in English and in Twitter in Spanish. We compare psycho-linguistics features in the suicide and non-suicide classes and also between languages. Regarding the results, in both languages we found that in suicide related messages more first singular person pronouns are used. In the English dataset the length of the messages is higher but not in the Spanish one due to the length restrictions in Twitter. The number of question-marks and the use of verb tenses are not significant to distinguish both classes.*
KEYWORDS: Language Processing, suicide, social network, language features

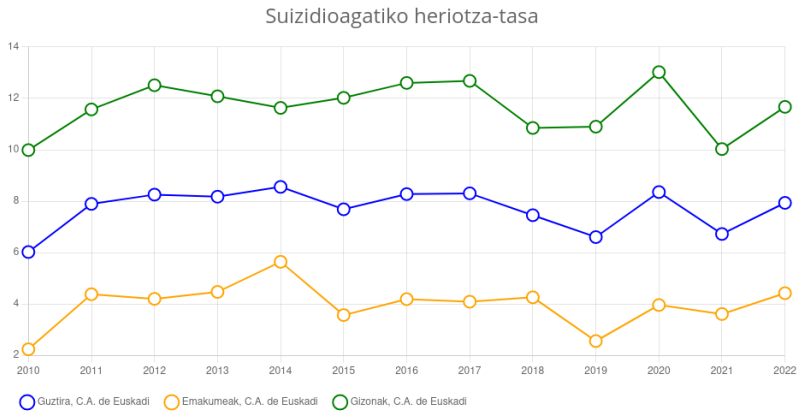
***Harremanetan jartzeko/Corresponding author:** Maite Oronoz, HiTZ Hizkuntza Teknologiako Euskal Zentroa - Ixa Taldea, Euskal Herriko Unibertsitatea (UPV/EHU), Informatika Fakultatea, Donostia-San Sebastián, Gipuzkoa.
https://orcid.org/0000-0001-9097-6047, maite.oronoz@ehu.eus

Nola aipatu/How to cite: (Adibidea:) 1. egilearen abizena, izena; 2. Egilearen abizena, izena...(202X). «Izenburua», Ekaia, DOI: https://doi.org/10.1387/ekaia.xxxxxx

Jasoa: xxxx, 202x; Onartua: xxxx, 202x
ISSN 0214-9001-eISSN 2444-3225 / ©2020 UPV/EHU

1. Sarrera

Suizidioa bakoitzak bere burua hiltzeko apropos egiten duen ekintza moduan definitzen da [1, 2]. Ekintza honen amaiera ez denean heriotza, suizidio-ahalegin kontzeptuaz ari gara. Osasunaren Mundu Erakundearen (OME) arabera [3] orduro 80 pertsonek egiten dute bere buruaz beste, eta suizidioak urtero 700.000 heriotza sortzen ditu. EUSTATen arabera, Euskal Autonomia Elkargoan eta 2022 urtean 100.000 pertsonako 7,94 pertsona hil ziren heriotza-arrazoia suizidioa izanik, gehienak gizonezkoak zirelarik. Suizidio-tasaren bilakaera ikus dezakegu 1.irudian 2010etik hasita 2022ra arte.



1. irudia: Suizidio-tasa EAEn

Pertsona bat bere buruaz beste egitera eramaten duten seinaleak identifikatzea ezinbestekoa da suizidio-arriskuan dagoela ezagutzeko eta ahal dela, neurriak hartzeko. Arrisku-faktoreen artean daude osasun mentalarekin zerikusia duten gaixotasunak, bakardadea edo hurbileko pertsonekin konexio falta, depresioa, substantzia-kontsumoa etab. Faktore zaurgarrien artean, osasun egoeraz gain, bestelakoak ere badaude: faktore sozioekonomikoak [4], lanbideak, adina (pertsona gazte [5] eta edadetuen artean [6] desberditasunak daude), generoa eta bestelako faktore sozial eta demografikoak.

Munduko Forum Ekonomikoak argitaraturiko “Global Governance Toolkit for Digital Mental Health” dokumentuan [7] diotenaren arabera "Teknologia disruptiboak osasun mentalean erabiltzeak irtenbide iraultzaileak sortzeko aukera ematen du osasun mentalean eta ongizatean emaitzak hobetzeko, inoiz baino neurri handiagoan."[1]. Adimen artifiziala erabiltzea portaera kolektiboetako ezagutza bereganatzeko ez da berria eta lehenago ere erabili izan da, adibidez politikaren gizartebehatokietan [8].

Artikulu honetan aurkeztutako lana OBSER-MENH izeneko ikerketa-proiektuan kokatzen da[2] eta honen helburua hizkuntza-teknologiak erabiliz sare sozialetan osasun mentalaren behatokia izatea da. Proiektuko Ixa taldeko kideok LOTU azpi-proiektuan[3] egiten dugu lan ezaugarri psikolinguistikoen azterketa eginez sare sozialetan ahalik eta azkarren bakardade eta isolamendua jasaten duten erabiltzailetan joera aldaketak detektatzeko hizkuntzaren ulermen sakona eta horretarako hizkuntza-teknologiak erabiliz.

Psiko-linguistikak ahozko portaeraren eta azpian dauden prozesu psikologikoen arteko harremanak aztertzen ditu. Lan honen kasuan, idatzizko testuetan (txioak, blogetako testuak, suizidio-oharrak, . . .) ezaugarri zehatz batzuk jorratzen dira aztertutako bibliografiaren arabera, suizidio-ideiagintzarekin zerikusia dutelakoan. Ezaugarri horien adibide dira mezuen luzera edota lehen

[1] *“Disruptive technology in mental health provides an opportunity to create breakthrough solutions that improve mental health and well-being outcomes on a greater scale than ever before”*

[2] OBSER-MENH proiektua: `http://nlp.uned.es/obser-menh-project/index.html`

[3] LOTU azpi-proiektua: `http://ixa.si.ehu.es/node/13593`

pertsonako izenordainen, aditz-denbora ezberdinen edo galdera-marken erabilera, esaterako. Artikulu honetan lan ezberdinetan identifikatutako ezaugarri horiek bildu eta bi izaera ezberdineko datu-sortetan aztertuko dira, ezaugarri bereizgarri horiek zein neurritan betetzen diren ezagutu nahian. Lana, esan bezala, sare sozialetako testuak analizatzera bideratu da.

Lehenago ere erabili izan dira sare sozialetan publikoki eskuragarri dauden datuak populazio mailako osasun mentala ebaluatzeko, [9] eta [10] lanetan, besteak beste. Osasun-erakundeetatik kanpoko datuen analisiak, sare sozialetako analisiak bereziki, suizidio arriskua detektatzeko sistemen garapena ahalbidetzen du, horretarako gizakiek bere sare sozialetan partekatzen dutena aztertuaz. Sare sozialetako informazioak maiztasun handiagoaz, eta batzutan egunero, pertsonen bizitza pribatuan barneratzea ahalbidetzen du eta modu honetan bere buruaz beste egiteko arriskuan dagoen pertsona bat modu eraginkorragoan beha daiteke [11].

Hurbilpen hauek erabili izan dira ingelesez idatzitako datu-bildumetan baina ez hainbeste gaztelaniaz idatzitako mezuetan, datu-bildumak urriagoak baitira. Azpimarratzekoa da, kasu honetan bezala, suizidioari buruzko azterketak automatikoki egiteko, gai honen inguruko datu-bildumak gaiari buruzko nahikoa bolumena izan behar duela. Hau horrela, oraingoz azterketa ingelesez eta gaztelaniaz egingo dugu. Gazteen Euskal Behatokiak 2022 urtean Euskararen Egunean euskararen ezagutzari eta sare sozialetan erabiltzeari buruz aurkeztutako lanaren arabera, Euskadiko gazteen $\%$ $36,1$ek diote euskara erabiltzen dutela sare sozialetan[4]. Joseba Fernandez de Landak [12] lanean euskal txiolariek idatzitako 6 milioi txio baino gehiago lortu zituen ia 8.000 erabiltzaile ezberdinenak. Orduan helburuetako bat helduek zein gai lantzen dituzten (politika, euskal presoak etab.) eta gazteek zeintzuk gairi buruz aritzen diren (bizitza kontatu, sentimenduak azalerazi etab.) aztertzea izan zen, baina etorkizunean corpus hau balia dezakegu suizidioaren inguruko azterketa bera egiteko. Oraingoz ingelesez egingo dugu azterketa dagoeneko sortua dagoen datu-bilduma batean eta gaztelaniarako, guk biltegiratu dugun datu-bilduma batean.

Laburtuz, lan honen **helburua** sare sozialetako testuetan agertzen diren eta suizidioarekin zerikusia duten ezaugarri psiko-linguistikoak aztertzea da, ezaugarri berberak ingelesez eta gaztelaniaz idatzitako testuetan aztertuaz.

Bi **ikerketa-galderei** erantzun nahi diegu:

IG1. Aurrekariek identifikatutako ezaugarri linguistikoak bereizten dira lan honetan aukeratutako datu-sortetan?

IG2. Ezaugarri linguistiko horiek desberdinak dira hizkuntza ezberdinetan?

Laburtuz, lan honetan suizidio-ideiagintza landuko dugu ingelesez eta gaztelaniaz idatzitako sare sozialetan, eta zehatz-mehatz honako ezaugarriak aztertuko ditugu: mezuen luzera, lehen pertsonako izenordainak, aditzen denbora eta galdera-marka kopurua.

2. Erlazionatutako lanak

Osasun mentalarekin arazoak dituzten pertsonek hizkuntzarekin zerikusia duten patroi bereizgarriak erabiltzen dituzte. Hizkuntzaren analisiaren bitartez eskizofrenia [13], nahasmendu obsesibo-konpultsiboa [14] eta mugako nortasunaren nahasmendua [15] identifikatu izan dituzte. Egile batzuk [16] lineako laguntza-taldeetatik informazioa erauzi izan dute, emozioen adierazpenak tarteko direlarik. [17] lanean erabiltzaileak osasun mentaleko hizkuntz markatzaileekin karakterizatu zituzten eta hizkuntzak bakardadea iragar dezakeen aztertu zuten.

Ideiagintza suizidari dagokionez, lan askok ondorioztatu dute lehen pertsona singularreko izenordain gehiago erabiltzen direla ideiagintza suizida presente dagoen poema [18], suizidio-ohar [19], txio [20, 21] eta Redditeko mezuetan [22]. Hainbat kasutan ikusi da testu hauek heriotzarekin

[4]Estatistikaren iturria: `https://www.euskadi.eus/eusko-jaurlaritza/-/albistea/2022/euskadiko-gazteen-36k-euskara-erabiltzen-du-sare-sozialetan/`

erlazionatutako hitz gehiago erabiltzen dituztela, edota honi erreferentzia gehiago egiten zaizkiola [18, 20]. Gainera, emozio negatiboekin lotutako hitzak ohikoagoak direnaren ebidentzia aurkitu da [19, 23], eta baita ezeztapena adierazteko hitz gehiago [23] eta indartzaile gehiago [21] erabiltzen direnarena ere, esandakoa biziagotzeko. Aditzei dagokienean, suizidio-oharretan geroaldiko aditz gutxiago erabiltzen direla uste da [19], eta txioetan orainaldian fokua jartzen dela [20].

Bestalde, suizidio-ideiagintzarekin erlazionatutako txioak luzeagoak direla [20] eta Redditeko mezuen kasuan galdera-marka gehiago erabiltzen dituztela [22] ikusi da.

Laburbilduz, aurrekarien arabera suizidio-ideiagintzaren zantzuak dituzten iturri ezberdinetako testuek hainbat ezaugarri linguistiko bereizgarri izan ohi dituzte. Lan horietan datu-bilduma ezberdinak erabili dira eta ez dago argi adierazpen horiek beste sorta batzuetara orokortu daitezkeen. Hauek dira aztertutako lanetan identifikatu diren ezaugarriak: lehen pertsona singularreko izenordain gehiago, heriotzarekin erlazionatutako hitz gehiago, emozio negatiboekin lotutako hitz gehiago, ezeztapena adierazteko hitz gehiago, indartzaile gehiago, geroaldiko aditz gutxiago, fokua orainaldian, mezu luzeagoak eta galdera-marka gehiago.

Eskuragarri dauden datu-bildumei dagokionez, ingelesez badaude batzuk eskuragarri, besteak beste, *Detection of Suicide Ideation in Twitter using Machine Learning* Twitterreko datu-sorta, *Suicide and Depression Detection* Reddit-ekoa eta *The University of Maryland Reddit Suicidality Dataset* [24] Reddit sare sozialekoa. Gaztelaniaz sare sozialetako testuekin osatutako bildumen artean anorexiarekin eta honen sintomak dituen txio bilduma [25]; osasun-mentalarekin erlazionatutako jokaera okerrak biltzen dituen Twitterreko bilduma [26]; eta azken aldian *MentalRiskES* lehiaketarekin lotutakoak daude. Azken txapelketa honetan 2023 urtean [27] elikadura-portaeraren nahasmenduekin, depresioarekin eta aurreko biak ez diren gaixotasun ezezagunekin lan egin zen. Aurten, 2024ean, depresioa/antsietatea eta suizidio-ideiagintza landuko dira.

3. Materialak eta Metodoak

3.1. Datuak jasotzea

Ingelesezko eta gaztelaniazko datu-bildumak izaeraz eta jasotzeko moduan, desberdinak dira. Ingeleserako Reddit sare-sozialetako mezuak erabili dira eta datu-bilduma dagoeneko sortuta zegoen. Xehetasunak eta deskribapen kuantitatiboa 3.1.1. atalean egin dugu. Gaztelaniarako guk geuk bildu dugu datu-sorta Twitterretik abiatuta eta hau egiteko metodologiaren xehetasunak 3.1.2. atalean emango ditugu.

3.1.1. Ingelesa

Ingeleserako, *Suicide and Depression Detection*[5] deritzon corpusa erabili da. Hau, Reddit sare sozialeko *"SuicideWatch"*, *"depression"* eta *"teenagers" subreddit* edo foroetatik eskuratutako mezu bilduma da. "SuicideWatch" foroko mezuak 2008/12/16tik 2021/01/02ra bitartekoak dira eta "depression" forokoak, aldiz, 2009/01/01-2021/01/02 tartekoak. Guztira, 232.704 mezu biltzen dira. Bildumaren deskribapenean ez da adierazten zenbat erabiltzailek idatzi dituzten mezuak.

Mezuek etiketa bana dute esleituta, alegia, klasea. Etiketa hori automatikoki esleitu da. Hau da, "*SuicideWatch*" foroaren helburua pentsamendu suizidak dituen orok laguntza eskatu ahal izateko espazio bat izatea denez, bertako mezuei *suizidioa* etiketa esleitu zaie eta "*teenagers*" foroko mezuei, aldiz, *ez-suizidioa*. *"depression"* azpimultzoko mezuak ez dira erabili lan honetan.

Klase banaketari dagokionean, banaketa uniformea da. Berez datu-sortak 1.taulako Corpusa zutabeko banaketa zuen. Hauek, ordea, ez dira lan honetan aztertu diren mezuak. Mezu hauek lehenik aurreprozesatu egin dira (azentu markak ezabatu, kontrakzioak zabaldu, URLak ezabatu

[5]Datu-sorta eskuragarri hemen: `https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch`

etab.) eta gero garbitu egin dira (testu hutsak ezabatu, mezu luzeak ezabatu etab.). Azkenean, 175.825 mezuekin egindako analisiak dira 4. atalekoak.

	Corpusa	Aurreproz.	Garbiketa
suizidioa	116.037 (% 50)	116.037 (% 50)	68.396 (% 38,9)
ez-suizidioa	116.037 (% 50)	116.037 (% 50)	107.429 (% 61,1)
mezu kopurua	232.074	232.074	175.825

1. taula: **Ingelesezko** datu-sortaren deskribapena. Datuak hiru zutabetan banatzen dira, bakoitza prestaketako fase bati dagokiolarik: Jatorrizko corpusa; aurreprozesaketa eta gero; garbiketa eta gero.

3.1.2. Gaztelania

Gaztelaniaz aipatutako ezaugarriak lantzen hasi ginenean, ez zeuden oraindik eskuragarri bibliografian aipatutako suizidioaren inguruko testu-bildumak. Hau horrela, corpusa gure kabuz jasotzeari ekin genion. Horretarako, honako metodologia jarraitu genuen:

- *Datu-iturria erabakitzea*. Bibliografiaren arabera Reddit eta Twitter dira datuak jasotzeko erabiltzen diren sare sozial ohikoenak. Reddit-en edukia orokorrean ingelesez dagoenez eta Twitter-en 2022-2023 urteetan 372,9 milioi aktibotatik Espainian 7,8 milioi eta Mexikon 11,8 milioi erabiltzaile aktibo egon zirenez [28], Twitter aukeratu genuen datu-iturri moduan.

- *Mezuak jasotzeko hitz-gakoak erabiltzea*. Hainbat aurrekarietan bezala [29, 30, 31] mezuak hitz-gakoak erabiliz jaso dira: i) afektibotasun-adierazleak [32], ii) aditu klinikoek aukeraturikoak [29], iii) depresioa tratatzeko *"inhibidores selectivos de la recaptación de serotonina (ISRS) ´´* botika klaseko botika-izenak [30] eta iv) Reddit-eko esaldi-gakoak [31]. Lehen bi motatakoak dira "*Deprimid (o/a/e)*[6]*, Cansad (o/a/e)...*" adibideen modukoak eta ISRS tratamenduen artean topatzen ditugu "*Prozac, Fluvoxamina ...*" eta antzekoak. Hitz-gakoak banan banan erabiltzera, bere buruaz beste egitearekin zerikusi gutxi duten mezuak jaso daitezkeenez mezuak irakurri ondoren 2.taulako moduko konbinaketak erabili dira (zerrenda luzeagoa den arren, bizpahiru adibide soilik erakutsi ditugu). Bilaketarako konbinazio hauekin mezuak eta erabiltzaileak eskuratu dira.

(quiero OR ojala OR deseo OR pienso) AND (morir OR vida OR suicidar OR (no AND vivir)) AND (sufrimiento OR "")
Estoy harta de este sufrimiento. No quiero vivir más.
(hospital AND suicid) OR (ingresad AND suicid)
El otro día intenté suicidarme y ahora estoy ingresado.
(intent OR trat) AND (suicid OR (quitar AND vid))
¿Y si intento quitarme la vida?

2. taula: Kontsulten zerrendarako eragile gutxi batzuk eta lortutako esaldien adibide bana.

- *Eskuzko sailkapena*. Mezuak irakurri egin ditugu erabiltzaileak *suizidioa* eta *ez-suizidioa* multzotan sailkatzeko. Eskuzko sailkapenean hiru multzo egin dira: i) suizidio arriskuan daudela esplizituki adierazi duten erabiltzaileak zeintzuen 10 mezu kasu positibotzat hartu diren (ikus 1 eta 2 adibideak 3.taulan), ii) suizidioarekin erlazionatutako gaiak aipatzen dituzten erabiltzaileak (aholkulariak edo bere buruaz beste egin duten pertsonen ezagunak eta

[6]Bilaketak *Deprimid* modukoekin egin dira, generoa eta zenbakia adierazten duten partikulen (-a, -o, -e, -as, -os, -es) erabilera ekiditeko.

(3. adibidea 3.taulan) iii) iragazki automatikoan jaso arren gaixotasun mentalarekin loturarik ez dutenak (4. adibidea 3.taulan). Lehen multzoko erabiltzaileak *suizidioa* klasean kokatu dira, azken bi multzotakoak *ez-suizidioa* klasean. 'Suizidio ideiagintza' eta 'ez-suizidioa' edo kontrol klasearen distribuzioa erabiltzaileetan uniformea (50:50 ratioan) izan dadin saiatu gara. Erabiltzaile bakoitzetik 60 egunetan bidalitako mezuen testua jaso dugu.

Mezuak jasotzeko prozesuan kontu handia izan dugu sailkapena ahalik eta hobekien egiteko, eta gainera, mezuak jasotzen aritu gineneko epean (2022 eta 2023 urteak) Twitter murriztapenak ipintzen hasi zen, beraz, datu-sorta hau txikiagoa da: guztira 91 erabiltzaile eta 11.776 mezu. Lortutako erabiltzaile eta mezu kopuruak 4.taulan topa daitezke.

	Mezua	Positiboa?
1	Ya tengo ganas de matarme otra vez.	BAI
2	Actualización de mi vida, me llevaron al hospital por intoxicación con medicamentos, vomité, me desmaye, me dieron arritimtias, me metieron una sonda por la nariz para hacerme un lavado gástrico y ahora llevo cuatro días en el hospital esperando a q me remitan a un psiquiátrico.	BAI
3	Hoy es el aniversario del suicidio de XXX. En su colegio no le dejaron encajar. Le hicieron sentir que no valía lo suficiente. Esto está judicializado, denunciado por sus padres y a la espera de juicio.	EZ
4	Prefiero suicidarme antes de besarme otra vez con ese tipo.	EZ
...	...	...

3. taula: Suizidioarekin zerikusia izan lezaketen mezu batzuk. Guztiak jaso dira modu automatikoan hitz-gakoak dituztelako, baina irakurri egin dira suizidio-ideiagintzan positiboak ote diren edo ez jakiteko (azken zutabea).

	Erabiltzaileak	Mezu kopurua
Suizidioa	47	6.986
Ez-suizidioa	44	4.790
Guztira	91	11.776

4. taula: **Gaztelaniako** datu-sortaren deskribapena: erabiltzaile eta mezu kopurua klaseka.

3.2. Tresnak

Izenordainak eta aditz-denborak aztertzeko analizatzaile linguistikoak behar dira. Kasu honetan, ikasketa sakonean oinarritutako UDPipe analizatzailea erabili da hitzen kategoria gramatikala eta ezaugarri linguistikoak jasotzeko.

UDPipe neurona-sare batez osatua dago eta esaldien segmentazioa, tokenizazioa, gramatika-kategoriaren (POS, *part-of-speech*) etiketatzea, lematizazioa eta mendekotasunetan oinarritutako analisi sintaktikoa egiteko gai da [33]. Gainera, *Universal Dependencies* (UD) proiektuan eraikitako *treebank* gehienetarako ereduak daude eskuragarri. Kasu honetan, ingeleserako eta gaztelaniarako prestatutako ereduak erabili dira.

UDPipe nazioarteko *Multilingual Parsing from Raw Text to Universal Dependencies* erronkaren testuinguruan sortu zen, CoNLL konferentziak antolatua [34]. Honengatik, eredua gai da analisia CoNLL-U formatuan adierazteko. Formatu honetan analizatutako hitz bakoitzeko informazio ugari lortzen da, hala nola, hitzaren lema, Universal POS (UPOS) eta hizkuntzaren menpekoa den POS (XPOS) etiketak, ezaugarri morfologikoak edota hitzen arteko dependentzia-erlazioak. Adibidez, ingelesezko *The* hitzak DET (determinatzailea) luke UPOS gramatika-kategoria unibertsal moduan eta DT (determinatzailea) ingelesezko treebank-ean ikasitako hizkuntzaren mendekoa den XPOS etiketa moduan.

UDPipek 2018ko *CoNLL* konferentzian emaitza onak lortu zituen[7], erabiltzeko oso erraza da, hainbat hizkuntzatarako eskuragarri dago eta behar dugun informazio linguistiko guztia ematen digu.

Lan honetan, zehazki, lehendabizi mezuak UDPiperekin tokenizatu egin dira eta ondoren, izenordainak eta aditz-denborak aztertzeko, token horien lemak, XPOS etiketak eta ezaugarri morfologiko batzuk izan dira kontuan.

Mezuen luzera aztertzeko, programazio-lengoaiek testuak hitzak espazioaren arabera banatzeko eskaintzen dituzten funtzioak erabili dira.

4. Emaitzak

Atal honetan bai ingelesez baita gaztelaniaz ere, suizidio eta ez-suizidio mezuetan honakoak aztertuko ditugu: mezuen luzera hitz kopurua kontuan hartuta (4.1. azpiatala), galdera-marken kopurua (4.2. azpiatalean), izenordainen erabilera (4.3. azpiatalean) eta aditz-denboren erabilera (4.4. azpiatalean). Garrantzitsua da gogoratzea, azterketa hauek bi testu-sorta zehatzetan egin direla eta gerta litekeela beste corpus batzuetan analisiaren emaitzak ezberdinak izatea.

Kasu guztietan datuak lehenengo taulen bitartez adieraziko dira eta gero kutxa-bibote diagramak erabiliz.

Tauletan mezuetan egindako azterketak bost estatistikoetan laburbildu dira: batezbestekoa ($\bar{x}$), desbiderapen estandarra (σ), minimoa (min), maximoa (max) eta hiru kuartilak: % 25a edo lehenengo kuartila, % 50a edo bigarren kuartila (medianaren modu berean kalkulatzen da) eta % 75a edo hirugarrena.

Luzeraren kasuan ezik, gainerako analisi guztietan datuak bi modutan eman dira: maiztasun absolutuak (abs. laburdura tauletan) eta luzerarekiko [0-100] eskalan normalizatuak (norm. laburdura tauletan) bereiziz. Datu normalizatuak emateak bi klaseen arteko konparaketa ahalik eta justuena egitea du helburu. Kutxa-bibote diagramak luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraiki dira.

4.1. Mezuen Luzera

Aurrekarietan aipatu da suizidio-ideiagintzarekin erlazioa duten testuak besteak baino luzeagoak izateko joera dutela [20]. Ezaugarri hori ingelesez Reddit sare-sozialeko testu-bilduma zehatz honetan eta gaztelaniaz Twitterreko mezu bilduma honetan ematen den aztertuko da atal honetan.

Ingelesa (5a taula eta 2a irudia). Klaseen arteko bereizketa oso argia dela ikus daiteke bai taulan, baita irudian laburbildutako emaitzetan ere. Suizidio-ideiagintzaren zantzuak erakusten dituzten mezuen batez besteko luzera 203 hitzekoa da, gainontzeko mezuena 61 hitzekoa den bitartean. Kuartilei dagokienean, ez-suizidioa etiketadun mezuen % 75a 60 hitzetik beherakoa da, ez-suizidioa kasuan 60 hitzetik beherako mezuen portzentaia % 25ekoa den bitartean. Hau oso garbi ikusten da diagraman, bi kutxak ez baitira puntu batean ere gainjartzen.

Gaztelania (5b taula eta 2b irudia). Kontuan hartu behar da Twitter sare sozialak duen ezaugarri bat: txioek gehienez 280 karaktere izan ditzakete. Beraz, luzeraren sakabanaketa beste sare sozialetan baino askoz txikiagoa izango da. Hain zuzen ere, suizidio ideiagintzarekin erlazionaturiko mezuen bataz besteko luzera 17,63 hitzekoa da eta ez-suizidio klasearena 17,01. Kuartiletan oso ikusgarria da datuak ia-ia berdinak direla. Alde handiena, medianan, hitz bakarrekoa da soilik. Gainera, bi klaseetan desbideratze estandarra parean dago (<0.2-ko diferentzia), hau da, bi multzoetan datuen sakabanaketa antzekoa da.

Konparaketa eta ondorioak

Ingelesezko datu-sortan mezuen luzera atributu iragarle gisa erabiliko balitz, 60 hitzetik gorako mezuei *suizidioa* etiketa esleituz eta aurkako kasuan *ez-suizidioa* klaseko mezutzat hartuz,

[7] https://universaldependencies.org/conll18/results-las.html

% 74,7ko asmatze-tasa lortuko litzateke. Honek, printzipioz, hasierako hipotesia baieztatzen du, Reddit sare sozialeko mezuak oinarri dituen datu-sorta honetan ideiagintza suizidarekin lotutako mezuak luzeagoak dira.

Gaztelaniako datu-sortan ohartu ahal da arrisku klaseko mezuak luzera maximoa handiagoa izateko tendentzia txikia dela. Oro har, bi klaseen arteko desberdintasuna ez da bereizgarria eta, beraz, mezuen luzeraren ezaugarria datu-sorta honetan ez dela esanguratsua esan daiteke.

Ingelesez Reddit sare-sozialean dirudienez mezuen luzera ez dago mugatua (maximoa 8.220 eta 9.684 hitzen artekoa da) eta Twitterren, ordea, muga dago (maximoak 75 eta 67 hitzen artekoa). Gure ustez, honek baldintza ditzake emaitza hauek.

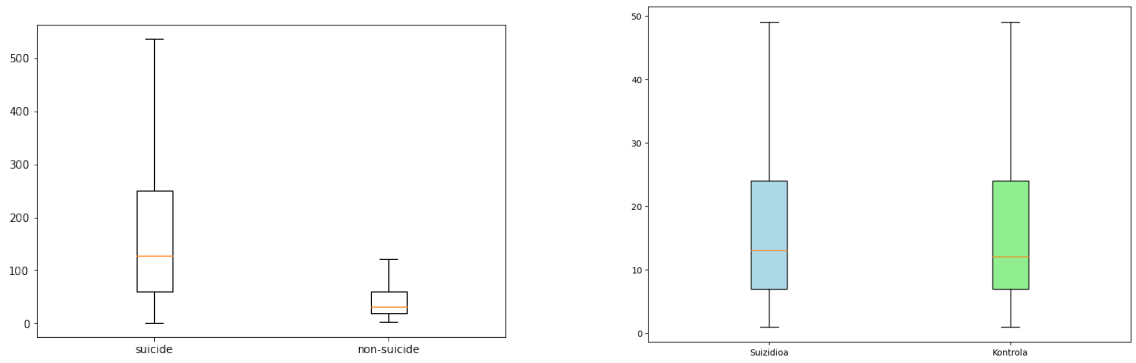
	suizidioa	*ez-suizidioa*
$\bar{x}$	203	61
σ	255	139
min	1	2
% 25	60	19
% 50	127	31
% 75	251	60
max	9.684	8.220

(a) Ingelesa.

	suizidioa	*ez-suizidioa*
$\bar{x}$	17,63	17,07
σ	13,76	13,92
min	1	1
% 25	7	7
% 50	13	12
% 75	24	24
max	75	67

(b) Gaztelania.

5. taula: Mezuen luzerarekin, hitz-kopuruarekin, erlazionatutako estatistikoak klaseka.



(a) Ingelesa. (b) Gaztelania.

2. irudia: Mezuen luzeraren kutxa-bibote diagrama klaseka.

4.2. Galdera-markak

Suizidio-ideiagintzarekin lotutako Redditeko mezuetan galdera-marken presentzia handiagoa denaren ebidentzia aurkitu da [22]. Hau proiektu honen ardatz diren bi datu-sorten gainean frogatzeko, mezu bakoitzean '¿' eta '?' ikur kopurua zenbatu da.

Ingelesa (6a taula eta 3a irudia). Taulan ikus daitekeenez *ez-suizidioa* etiketadun mezuen kasuan handiagoak dira batezbesteko eta desbideratze estandar normalizatuak. Hala ere, bi etiketen kasuan mediana 0 da, hau da, bi klaseetako mezuen erdiak ez du galdera-markarik erabiltzen. Kutxa-bibote diagramari erreparatuz esan daiteke suizidio-ideiagintzarekin erlaziorik ez duten mezuen kasuan aldakortasuna handiagoa dela. Maiztasun absolutuak aztertuz, bi kasuetan mezuek galdera-marka bat baino gutxiago erabiltzen dute.

Aipatu beharrekoa da *ez-suizidioa* etiketadun mezu batek 8.209 galdera-marka erabiltzen dituela. Mezu hau aztertuta ikusi da salbuespen bat dela, funtsean galdera-marka segida luze batek osatua. Segida horrek esanahi berezirik ez duenez, kasu isolatutzat jo da, alde batera utziz.

Gaztelania (6b taula eta 3b irudia). Begirada azkar bat botatzea nahikoa da argi ikusteko bi klaseetako mezu gehienek ez dutela galdera-markarik erabiltzen. Hain zuzen, mezuen % 75ak baino gehiagok ez du galdera-markarik erabiltzen.

Ikur gehien erabiltzen dituzten mezuek 12 galdera-marka erabiltzen dituzte bi klaseetan. Gainera, datu normalizatuei erreparatuz ikusten da kasu maximoak bereziak direla: bi klaseetan galdera-ikurrak mezuaren % 66a baitira, txio hauek *"Qué?????????????"* edota *"Qué dices?????????????"* motakoak dira.

Konparaketa eta ondorioak

Lan honetan ingeleserako erabilitako datu-sortaren kasuan ezin daiteke esan suizidioarekin erlazionatutako mezuek gainontzekoek baino galdera-marka gehiago erabiltzen dituztenik. Are gehiago, nahiz eta orokorrean mezu guztietan ikur hauen presentzia txikia den, berez *ez-suizidioa* gisa etiketatutako mezuek gehiago erabiltzen dituztela ikusi da.

Gaztelaniazko datuen azterketari dagokionez, mezu gehienek ez dute galdera-markarik erabiltzen. Galdera-markak analizatu dituzten aurrekariek testu luzeagoak baimentzen dituzten sare sozialekin lan egin dute, beraz, lan honetan eta aurrekarien arteko emaitza hain ezberdinak egoteko azalpen posible bat da. Esaterako, [22] lanean Redditeko mezuak analizatu zituzten, non 40.000 karaktereko mezuak idaztea posiblea da, Twitter baino 140 aldiz gehiago.

Dena den, argi dago ingelesez eta gaztelaniaz aztertutako datu-sortetan ez dela ia galdera-markarik erabiltzen. Galdera-marken erabilera ez da bereizgarria eta, beraz, ezaugarri hau ez da esanguratsua bi klaseak ezberdintzeko.

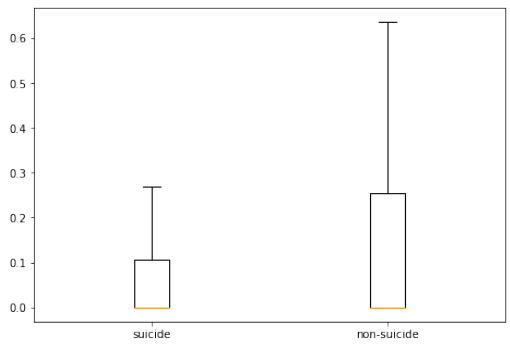
	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	0,76	0,13	0,70	0,27
σ	1,81	0,40	32,9	0,96
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	0	0,00	0	0,00
% 75	1	0,11	1	0,26
max	222	32,0	8.209	98,7

(a) Ingelesa.

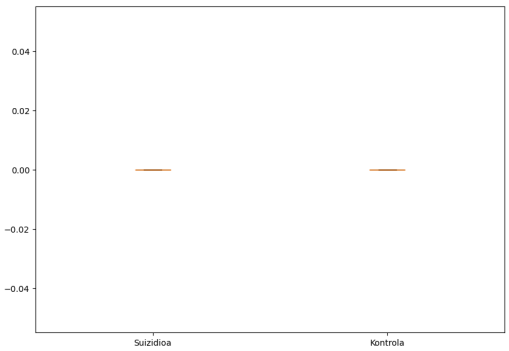
	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	0,12	0,73	0,15	0,89
σ	0,53	3,62	0,69	3,97
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	0	0,00	0	0,00
% 75	0	0,00	0	0,00
max	12	66,66	12	66,66

(b) Gaztelania.

6. taula: Galdera-markak. Mezuen galdera-marka kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.



(a) Ingelesa.



(b) Gaztelania.

3. irudia: Galdera-markak. Mezuen galdera-marka kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.

4.3. Lehen pertsona singularreko izenordainak

Hainbat lanek ondorioztatu dute suizidio-ideiagintzaren zantzuak erakusten dituzten testu mota ezberdinetan ohi baino lehen pertsona singularreko izenordain gehiago erabiltzen direla [18, 19, 20, 21, 22]. Ezaugarri linguistiko hau aztertzeko, UDPipe tresnaren (informazio gehiago 3.2. atalean) XPOS (hizkuntzaren menpeko *part-of-speech*) gramatika-kategoriaren etiketatzailea erabili da. Bai ingelesez, baita gaztelaniaz ere, izenordainaren mota adierazten duen etiketak erabili dira, pertsona-izenordainak eta edutezkoak, hain zuzen ere. Zehazki, PRP (*Personal pronoun*, pertsona-izenordain) eta PRP$ (*Possessive pronoun*, edutezko izenordain) gisa etiketatutako hitzak bilatu dira. Izenordainen pertsona eta zenbakia ezagutzeko ezaugarri morfologikoak aztertu dira, lehen pertsona singularrekoak direla bermatzeko.

Ingelesa (7a taula eta 4a irudia). Kasu honetan, bi etiketen arteko ezberdintasuna nabaria da maiztasun absolutuko datuak behatuz. Hirugarren kuartilari erreparatuz, esaterako, ikus daiteke *ez-suizidioa* etiketadun mezuen laurden batek soilik erabiltzen dituela 5 izenordain baino gehiago, *suizidioa* etiketadunen hiru laurdenek 7 baino gehiago erabiltzen dituzten bitartean. Normalizatutako datuen kasuan, nahiz eta ezberdintasun hori presente dagoen, ez da hain nabaria. *suizidioa* etiketadun mezuen kasuan lehen pertsona singularreko izenordainek batez beste hitzen % 10,7a hartzen dute, *ez-suizidioa* etiketadunen kasuan hitz hauen batez besteko portzentaia % 7,39koa den bitartean. Diagrama behatuz ikusten da suizidio-ideiagintzarekin erlazionatutako mezuetan izenordain hauen erabilera nabarmenagoa dela, medianaren diferentzia ia lau puntukoa izanik. Hala ere, *ez-suizidioa* kasuan bai kutxa eta bai biboteak luzeagoak dira, hau da, datuek sakabanatze handiagoa erakusten dute. Honek, suizidio-ideiagintzaren zantzurik ez duten mezuetan izenordain hauen erabilerak aldakortasun handiagoa duela esan nahi du.

Gaztelania (7b taula eta 4b irudia). Kasu honetan ere, kutxa-bibote diagrama begiratzen badugu 4b bi klaseen arteko ezberdintasunak nabariagoak dira. Taulari begiratzen badiogu, bigarren kuartilari (medianari) erreparatuz gero ohartu ahal da mezu gehienek ez dutela lehenengo pertsona singularreko izenordainik erabiltzen. Hala ere, badaude klaseen arteko ezberdintasunak eta hauek hirugarren kuartiletik aurrera nabaritzen dira. Erabilera maximoa begiratzen badugu ikus dezakegu, suizidio ideiagintza zantzuak dituzten erabiltzaileek gehienez 25 izenordain erabili dutela eta, aldiz, ez-suizidioa klasean dauden erabiltzaileek 7 izenordain. Datu hauek ikusi eta gero, ikerketa zehatzago bat egin da kopuru absolutuarekin: 3 izenordain edo gehiago erabiltzen dituzten mezuak bilatu eta zenbatu egin dira bi klasetan. Emaitzak hurrengoak izan dira: suizidio ideiagintzako zantzuak dituzten mezuen % 13,36ak hiru izenordain edo gehiago dituzte eta kontrol edo ez-suizidio klasean, ordez, soilik % 3,31ak dituzte 3 izenordain baino gehiago.

Kutxa-bibote diagramari begiratuta ikus dezakegu suizidio klaseko kutxa ez-suizidioa klasekoa baino askoz handiagoa dela, eta beraz, mezuen artean aldakortasun handiagoa dagoela.

Konparaketa eta ondorioak

Ingelesari dagokionez, mezuen luzera aintzat hartuta, esan daiteke *Suicide and Depression Detection* datu-sortan ere suizidio-ideiagintzarekin erlazioa duten mezuetan lehen pertsona singularreko izenordain gehiago erabiltzen direla gai horrekin erlaziorik ez duten mezuetan baino.

Gaztelaniakoetan ere, eta zenbaki absolutuak kontuan hartuta, esan daiteke lan honetan lortutako datu sortan suizidio-ideiagintzarekin erlazioa duten mezuetan lehen pertsona singularreko izenordain gehiago erabiltzen dutela.

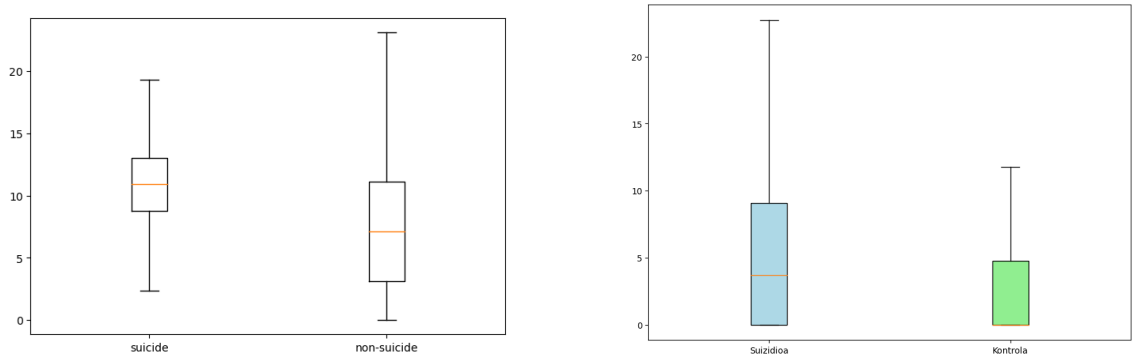
	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	24,6	10,7	4,90	7,39
σ	29,9	3,97	9,54	5,49
min	0	0,00	0	0,00
% 25	7	8,74	1	3,13
% 50	16	10,9	2	7,14
% 75	32	13,0	5	11,1
max	1.937	100,0	696	64,2

(a) Ingelesa

	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	1,08	5,47	0,47	3,22
σ	1,42	6,86	0,82	6,24
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	1	3,70	0	0,00
% 75	2	9,09	1	4,76
max	25	66,66	7	50,0

(b) Gaztelania.

7. taula: Izenordainak. Mezuen lehen pertsona singularreko izenordain kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.



(a) Ingelesa. (b) Gaztelania.

4. irudia: Izenordainak. Mezuen lehen pertsona singularreko izenordain kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.

4.4. Aditz-denborak

Hainbat lanetan, suizidio-ideiagintzarekin erlazionatutako testuetan denborarekin lotutako ezaugarriak identifikatu dira. Horietako batean [19], esaterako, suizidio-oharretan geroaldiko aditz gutxiago erabiltzen direla ondorioztatu zen, eta txioak aztertzen zituen beste batean [20] ideiagintza suizida presente dagoen txioetan orainaldian jartzen dela fokua.

Modu ezberdinak daude denbora hitzez islatzeko, denbora-aditzondoak edota aditzak kasu. Proiektu honetan bigarren hauetan jarri da fokua, eta lehenaldiko, orainaldiko eta geroaldiko aditzen erabilera aztertu da, bakarka lehendabizi, eta hirurak batera ondoren.

Ingelesez eta gaztelaniaz modu ezberdinean baina konparagarrian erauzi dira hizkuntza-ezaugarriak. Ingelesez XPOS etiketari begiratu zaion bitartean, gazteleraz ezaugarri morfologikoei begiratu zaie (aditza kategoriako *Tense*, *grammatical tense* edo denbora ezaugarriari, hain zuzen ere).

4.4.1. Lehenaldia

Lehenaldiko aditzak identifikatzeko XPOS etiketaren VBD balioa bilatu da (*Verb, past tense*, aditza lehenaldian) eta gazteleraz lehenaldiko markak bilatu dira: *"Tense=Past"* (lehenaldia) eta *"Tense=Imp"* (lehenaldi inperfektua). Ondoren, aditzon kopuruak gorde egin dira analisia egiteko.

Ingelesa (8a taula eta 5a irudia). Taulan eta diagraman ikus daiteke bi klaseetan lehenaldiko aditzen antzeko erabilera egiten dela. Ezberdintasun nagusia medianan dago, *suizidioa* etiketadun

mezuen kasuan handiagoa baita.

Gaztelania (8b taula eta 5b irudia). Emaitzak laburbiltzen dituzten taulan eta diagraman ikus daiteke bi klaseetan lehenaldiko aditzak ez direla gehiegi erabiltzen. Arrisku klaseko mezuak lehenaldia erabiltzeko tendentzia handiagoa dutela esan daiteke, baina izenordainekin konparatuz, ez da ezaugarri bereizgarria.

Aipatzekoa da ez-suizidioa klaseko datu normalizatuen maximoa 100 dela. Datu-sortan bilatu dira 100 aditz lehenaldian dituzten mezuak eta 4 kasu zehatz direla ikusi da. Horietako bat UD-Pipek egindako errore bategatik sortu da: erabiltzaileak *"Se acabó"* mezua idatzi nahian *"Sacabó"* (tarterik gabe) idatzi zuen. Hau horrela, programak *"Sacabó"* hitzak *"Sacabar"* aditz faltsuaren jokabidea duela kalkulatu du.

Konparaketa eta ondorioak

Badirudi bai ingelesezko baita gaztelerazko datu-sortetan lehenaldiko aditzak era berean erabiltzen direla suizidio-ideiagintzarekin erlazionatutako mezuetan eta honekin zerikusirik ez dutenetan.

	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	7,12	2,28	2,10	2,48
σ	16,2	2,46	6,72	3,65
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	2	1,63	0	0,00
% 75	7	3,45	2	4,14
max	590	37,5	643	42,9

(a) Ingelesa

	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	0,62	2,66	0,45	2,27
σ	1,26	5,12	0,92	5,45
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	0	0,00	0	0,00
% 75	1	4,00	1	2,32
max	12	50,0	8	100,0

(b) Gaztelania.

8. taula: Lehenaldia. Mezuen lehenaldiko aditz kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.

4.4.2. Orainaldia

Orainaldiko aditzekin lehenaldikoekin egin den antzera jokatu da, ingelesez XPOS etiketak erabiliz. Kasu honetan, bi etiketa daude orainaldiko aditzentzat: VBP (*Verb, non-3rd person singular present*), hirugarren pertsona singularrari erreferentzia egiten ez dioten orainaldiko aditzentzat eta VBZ (*Verb, 3rd person singular present*), hirugarren pertsona singularreko orainaldiko aditzei dagokiena. Gazteleraz *"Tense=Pres"* ezaugarria duten hitzak bilatu dira.

Ingelesa (9a taula eta 5a irudia). Luzerarekiko normalizatutako datuetan ageri den ezberdintasun nagusia datuen dispertsioan dago. Hau handiagoa da *ez-suizidioa* etiketadun mezuen kasuan, diagramako kutxetan ikusten denez. Suizidio-ideiagintzarekin erlazionatutako mezuen kasuan hirugarren eta lehenengo kuartilen arteko diferentzia 1,96 puntukoa da, gainontzeko mezuetan 6,34 puntukoa den bitartean. Ezberdintasun hau datuen desbideratze estandarrean ere beha daiteke. Normalizatutako mediana eta batezbestekoa, aldiz, oso antzekoak dira bi kasuetan.

Gaztelania (9b taula eta 5b irudia). Emaitzak behatuz orain arte bi klaseen arteko ezberdintasun gutxien duen ezaugarria da. Maiztasun absolutu baita luzerarekiko normalizatutako datuetan parekotasun handia ikus daiteke suizidio ideiagintza eta ez-suizidioa klaseen artean. Alde handiena datu normalizatuetako desbideratze tipikoan dago: arrisku klasearena kontrolarena baino puntu bat txikiagoa da.

Konparaketa eta ondorioak

Bi kasuetan, bai ingelesez baita gaztelaniaz ere, esan daiteke orainaldiko aditzen antzeko erabilera egiten dela, *ez-suizidioa* etiketadunen kasuan aldakortasuna handiagoa dela kontuan izanik.

Atalaren hasieran aipatu dira [19] eta [35]ek egindako lanak, non ondorioztatzen zuten suizidio ideiagintza-zantzuak dituzten testuek orainaldia erabiltzeko joera handiago zutela zantzurik gabekoek baino. Bi datu-sorta hauen kasuan, ezaugarri hura ez da betetzen, orain arte aztertutako ezaugarrien artean parekotasun handiena duenetarikoa baita.

	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	18,6	8,74	5,09	8,13
σ	20,2	3,40	11,1	5,04
min	0	0,00	0	0,00
% 25	6	6,67	1	4,76
% 50	13	8,64	3	7,87
% 75	24	10,6	6	11,1
max	848	50,0	1.263	50,0

(a) Ingelesa.

	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	1,69	9,03	1,52	8,74
σ	1,75	8,81	1,62	9,77
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	1	8,33	1	7,69
% 75	2	13,63	2	12,5
max	11	100,0	18	100,0

(b) Gaztelania.

9. taula: Orainaldia. Mezuen orainaldiko aditz kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.

4.4.3. Geroaldia

Geroaldiko aditzen kasua berezia da. Izan ere, ingelesezko aditzetan ez da geroaldia esplizituki adierazten, kontzeptu lausoa da. Askotan "*will*" hitza erabili ohi da horretarako. Hala ere, gai honen inguruko eztabaida dago literaturan, batzuek diotelako modala dela, ez geroaldiaren zuzeneko adierazgarri, eta geroaldia errepresentatzeko gaitasuna hitz honen modalitatetik datorrela [36]. Etorkizunari erreferentzia egiteko erabili ohi den beste egitura bat "*be going to*" da. Ikusten denez, gai konplexua da geroaldiarena ingelesez. Lan honetarako, aipatutako bi egiturak izan dira kontuan, "*will*" eta "*be going to*". Lehenengoaren kasuan, lema gisa "*will*" hitza duten tokenak bilatu dira, MD (*modal*) XPOS etiketa dutenak, eta dependentziei dagokienez, *head* edo burutzat aditz bat dutenak. Bigarrenari dagokionez, honakoa egin da "*be going to*" egitura identifikatzeko: "*be*" lemadun hitzak aurkitu dira, buru gisa "*going*" hitza dutenak. Ondoren, "*going*" hitz horren hurrengoa "*to*" den konprobatu da, eta azken honen buru den tokena aditza den, hau da, XPOS etiketa VB- (*Verb, ...*) karaktereez hasten den.

Gazteleraz *"Tense=Fut"* ezaugarria bilatu da. Gazteleraz "*viviré, leeremos...*" modukoetan analisiko ezaugarrietan adierazten baita geroaldiko denbora, ingelesez "*I will live, I am going to live*" edo "*we will sing, we are going to sing*" diogun bitartean (biziko naiz, abestuko dugu, hurrenez hurren).

Ingelesa (10a taula eta 5a irudia). Luzerarekiko [0-100] eskalan normalizatutako batezbesteko eta desbideratze estandarrari erreparatuta esan daiteke bi klaseek aditz-denbora honen antzeko erabilera egiten dutela, eta baita presentzia gutxi dutela ere. Maiztasun absolutuko datuei dagokienean, *suizidioa* klaseko mezuetan batez beste etorkizunari erreferentzia egiten dion aditz bat erabiltzen dela ikusten da, *ez-suizidioa* klasekoetan kopuru hori zerora hurbiltzen den bitartean. Diagraman, *ez-suizidioa* klasearen kasuan aztertutako egituren erabilera hutsala dela ikus daiteke. Suizidio-ideiagintza erakusten duten mezuetan, aldiz, aditz-denbora hau presente dago, nahiz eta kopuru txikian izan.

Gaztelania (10b taula eta 5b irudia). Gaztelaniaz ere oso txikia da geroaldian dauden aditzen agerpen kopurua. Suizidio ideiagintza klaseko mezuen % 3,27ak soilik izan dute geroaldian idatzitako aditzen bat. Ez-suizidioa kasuan zenbakiak txikiagoak dira, % 3,04koak izanik. Gainera, bi klaseen arteko maiztasuna nahiko parean dagoenez, ezaugarri ez esanguratsu bat dela esan daiteke.

Konparaketa eta ondorioak

Esan daiteke bi klaseetan etorkizunari erreferentzia gutxi egiten zaiola, aztertutako egiturei dagokienean behintzat. Bi hizkuntzetan kuartilei erreparatuz, guztiak zeroz osatuta daude.

	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	1,03	0,52	0,23	0,33
σ	1,70	0,96	0,80	1,12
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	0	0,00	0	0,00
% 75	1	0,69	0	0,00
max	57	20,0	68	33,3

(a) Ingelesa.

	suizidioa		*ez-suizidioa*	
	abs.	norm.	abs.	norm.
$\bar{x}$	0,04	0,19	0,03	0,13
σ	0,22	1,41	0,20	0,96
min	0	0,00	0	0,00
% 25	0	0,00	0	0,00
% 50	0	0,00	0	0,00
% 75	0	0,00	0	0,00
max	3	50,0	3	33,33

(b) Gaztelania.

10. taula: Geroaldia. Mezuen geroaldiko aditz kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.

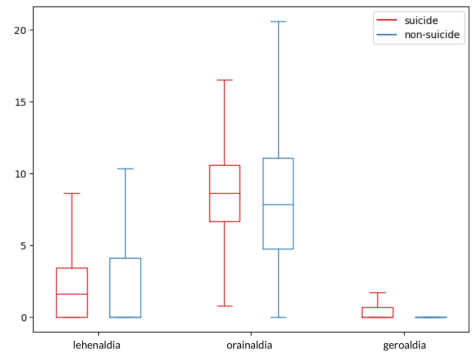
4.4.4. Hiru aditz-denboren arteko konparaketa

Hiru aditz-denborak era isolatuan aztertu ostean, beraien arteko konparaketa egitea erabaki da. Horretarako, 5a eta 5b irudietako kutxa-bibote diagramak eraiki dira, non bi hizkuntzetan aditz-denbora bakoitzari dagokion grafikoa irudikatu den bi klaseak koloretan bereiziz eta luzerarekiko [0-100] eskalan normalizatutako datuak erabiliz.

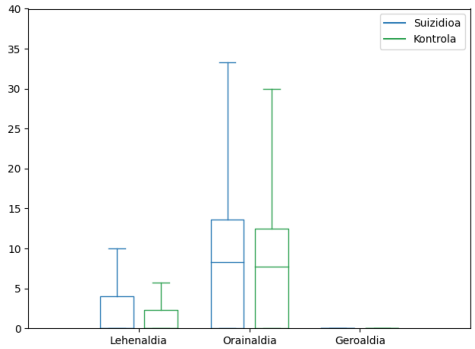
Ingelesez zein gaztelaniaz, bi klaseetan diferentzia esanguratsuarekin gailentzen den aditz-denbora orainaldia da, honi dagozkion kutxak ez baitira gainjartzen beste bi bikoteekin. Beraz, esan daiteke kasu honetan ere suizidio-ideiagintzarekin zerikusia duten mezuek orainaldian jartzen dutela fokua. Hala ere, gauza bera gertatzen da *ez-suizidioa* etiketadun mezuen kasuan eta ondorioz, hau ezin da suizidio-ideiagintzarekin erlazionatutako mezuen ezaugarri bereizgarritzat hartu.

Bestalde, etorkizunari erreferentzia egiten dioten geroaldiko aditzei dagokienean, ikusten da presentzia gutxiena dutenak direla bi kasuetan. Gaztelaniaz, geroaldiko aditzen presentzia nulua da bi klasetan. Beraz, oro har, esan daiteke Twittereko erabiltzaileek geroaldia erabiltzeko tendentziarik ez dutela, datu-bilduma honetan, behintzat. Beraz, emaitza hauek ikusita, ezin daiteke esan ideiagintza suizidaren zantzuak dituzten mezuek etorkizunari erreferentzia gutxiago egiten diotenik.

Laburbilduz, lan honetan lortutako bi datu-sortetan ez dira betetzen aurrekarietan aditz-denborekin erlazionatuta dauden eta aztertu diren bi ondorioak era argi batean.



(a) Ingelesa.



(b) Gaztelania.

5. irudia: Aditzak batera. Mezuen aditz-denboren erabileraren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.

5. Ondorioak eta etorkizuneko lana

Lan honetan beste egile batzuek aztertutako lau ezaugarri landu dira bi datu-sorten gainean: mezuen luzera, galdera-markak, lehen pertsona singularreko izenordainak eta aditz-denborak. Analisia bi hizkuntzetan egin da eta bi sare sozialetan. Ingeleserako, Reddit sare sozialeko *Suicide and Depression Detection* deritzon corpusa erabili da eta gaztelaniarako Twitterreko corpus bat lortu da.

Hona hemen bi hizkuntzetan aztertutako ezaugarrien konparaketa:

Mezuen luzera: Ingelesez idatzitako datu-sortan ondorioztatu da suizidio ideiagintzaren zantzuak erakusten dituzten mezuak luzeagoak direla. Gaztelaniakoan, bi klaseen arteko mezuen luzera parean dagoela ikusi dugu. Aipatu den bezala, arrazoia Twitterreko txioen 280 karaktereko mugan egon daiteke.

Galdera-marka kopurua: Galdera-markekin ez da esperotakoa bete. Ezaugarri honen kasuan bi hizkuntzetan ondorioztatu da bi klaseek antzeko maiztasunarekin erabiltzen dituztela galdera-markak. Beraz, ezaugarria ez dela nabarmentzekoa ondorioztatu da.

Lehen pertsona singularreko izenordainen kopurua: Aurrekarietan adierazi moduan Redditeko datu-sortan suizidio klaseko mezuek lehen pertsona singularreko izenordain gehiago erabiltzen dituzte. Gaztelaniaz ere betetzen da ezaugarri hau, nahiz eta aldea txikiagoa den bi klasetan.

Aditz-denboren erabilera: Aditz- denboren erabileran bi datu-sortetan aurkitu dira pareko emaitzak: bi klasetan gehien erabiltzen den aditz-denbora orainaldia da, baina ezin da hartu ezaugarri bereizgarritzat, eta geroaldiaren presentzia minimoa da. Lehenaldiari dagokionez, era berean erabiltzen dira aditzok suizidio-ideiagintzarekin erlazionatutako mezuetan eta honekin zerikusirik ez dutenetan.

Beraz, proiektu honetan mahaigaineratutako lehen ikerketa-galderari (IG1) erantzunez, lan honetan aukeratutako datu-sortetan aurrekarietan identifikatutako ezaugarrietako batzuk bereizten dira, ez denak, ordea.

Bigarren ikerketa-galdera (IG2) erantzuteko esango dugu luzeran badagoela aldea baina sare sozialak baldintzatutako dela, Twitterren luzera muga baitago. Lehen pertsona singularreko izenordain kopuruari dagokionez, ingelesez neurri handiagoan betetzen da suizidio-zantzudun klasean mota honetako izenordain gehiago erabiltzea. Gainerakoetan ez dago bi hizkuntzetan alde handirik.

Esan beharra dago, ondorio hauek jasotako datu-bildumen mendekoak direla eta mezuak jasotzeko moduak ere eragina izan dezakeela. Ingelesez, mezuak bi *subreddit* ezberdinetatik eskuratuak izan dira eta bakoitzaren jatorriaren arabera esleitu zaie etiketa, suizidio-ideiagintzan aditua den profesional batek prozesuan parte hartu gabe. Beraz, baliteke foro bakoitzaren izaeran eta bakoitzari ematen zaion erabileran ezberdintasunak egotea eta hau izatea aztertutako ezaugarrietan topatutako bereizketen arrazoia. Gaztelaniaz, prozesua zehatzagoa izanagatik, txio kopurua txikiagoa da eta orokortzea zailagoa da.

Lanaren mugak aipatu beharko bagenitu, esan beharra dugu erabilitako tresna, UDPipe ez dagoela sare sozialetan erabiltzen den hizkuntzara egokituta. Hori horrela akatsak egingo ditu ("*tqm - te quiero mucho*", "*xq - por qué*", etab.) modukoak analizatzean.

Analisi linguistikoari dagokionez, proiektu honetan hasitako analisiak ateak ireki dizkio abstrakzio maila altuagoko ezaugarriak aztertzeari. Izan ere, aurrekariek suizidio-zantzuak erakusten dituzten mezuetan ezeztapena presenteago dagoenaren edota emozio negatibo eta heriotzarekin erlazionatutako hitz gehiago erabiltzen direnaren ebidentzia aurkitu dute, baita indartzaileen (oso, guztiz, zinez...) presentzia nabarmenagoa denarena ere. Ezaugarri hauek analisian sakontzea eskatzen dute eta horretan gabiltza.

Bibliografia

[1] UNIVERSITY OF MANITOBA, 2021, «Concept: Suicide and attempted suicide (intentional self inflicted injury)», `http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1183`, [Accessed: 06/02/2024].

[2] P. MARTENS, R. FRANSOO, N. MCKEEN *eta kolaboratzaileak*, 2004, «Patterns of regional mental illness disorder diagnoses and service use in manitoba: a population-based study. winnipeg: Manitoba centre for health policy, 2004», .

[3] W. H. ORGANIZATION, 2023, «Suicide», `https://www.who.int/news-room/fact-sheets/detail/suicide`, [Accessed: 06/02/2024].

[4] F. FERRETTI eta A. COLUCCIA, 2009, «Socio-economic factors and suicide rates in European Union countries», *Legal Medicine*, **11**, S92–S94.

[5] D. WASSERMAN, Q. CHENG eta G.-X. JIANG, 2005, «Global suicide rates among young people aged 15-19», *World psychiatry*, **4**(2), 114.

[6] M. WAERN, E. RUBENOWITZ, B. RUNESON, I. SKOOG, K. WILHELMSON eta P. ALLEBECK, 2002, «Burden of illness and suicide in elderly people: case-control study», *Bmj*, **324**(7350), 1355.

[7] WORLD ECONOMIC FORUM eta DELOITTE, 2021, «Global governance toolkit for digital mental health: Building trust in disruptive technology for mental health», .
URL `https://www3.weforum.org/docs/WEF_Global_Governance_Toolkit_for_Digital_Mental_Health_2021.pdf`

[8] S. CATON, M. HALL eta C. WEINHARDT, 2015, «How do politicians use Facebook? An applied social observatory», *Big Data & Society*, **2**(2), 2053951715612822.

[9] A. FINE, P. CRUTCHLEY, J. BLASE, J. CARROLL eta G. COPPERSMITH, 2020, «Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data», *Proceedings of the fourth workshop on natural language processing and computational social science*, Orrialdeak 50–54.

[10] H. MORADIAN, M. A. LAU, A. MIKI, E. D. KLONSKY eta A. L. CHAPMAN, 2022, «Identifying suicide ideation in mental health application posts: A random forest algorithm», *Death Studies*, Orrialdeak 1–9.

[11] P. RESNIK, A. FOREMAN, M. KUCHUK, K. MUSACCHIO SCHAFER eta B. PINKHAM, 2021, «Naturally occurring language as a source of evidence in suicide prevention», *Suicide and Life-Threatening Behavior*, **51**(1), 88–96.

[12] J. F. DE LANDA, 2019, «Gazteak eta euskarasare sozialetan. zer, nori, nork: euskarazko txio formaleta informalak sailkatuz eta konparatuz», *Eusko Ikaskuntzaren XVIII Kongresua"Geroa Elkar-Ekina": mendeurreneko kongresua= XVIII Congreso de Estudios Vascos"El futuro que nos (re) une": congreso del centenario= XVIIIe Congrès d'Etudes Basques"Notre futur ensembre": 2018, Baiona, Vitoria-Gasteiz*, Orrialdeak 348–355, Sociedad de Estudios Vascos= Eusko Ikaskuntza.

[13] K. S. MINOR, K. A. BONFILS, L. LUTHER, R. L. FIRMIN, M. KUKLA, V. R. MACLAIN, B. BUCK, P. H. LYSAKER eta M. P. SALYERS, 2015, «Lexical analysis in schizophrenia: how emotion and social word use informs our understanding of clinical presentation», *Journal of psychiatric research*, **64**, 74–78.

[14] E. OREN, N. FRIEDMANN eta R. DAR, 2016, «Things happen: Individuals with high obsessive–compulsive tendencies omit agency in their spoken language», *Consciousness and cognition*, **42**, 125–134.

[15] P. E. CARTER eta B. F. GRENYER, 2012, «Expressive language disturbance in borderline personality disorder in response to emotional autobiographical stimuli», *Journal of personality disorders*, **26**(3), 305–321.

[16] B. B. C. DA SILVA eta I. PARABONI, 2018, «Personality recognition from facebook text», *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, Orrialdeak 107–114, Springer.

[17] S. C. GUNTUKU, R. SCHNEIDER, A. PELULLO, J. YOUNG, V. WONG, L. UNGAR, D. POLSKY, K. G. VOLPP eta R. MERCHANT, 2019, «Studying expressions of loneliness in individuals using twitter: an observational study», *BMJ open*, **9**(11), e030355.

[18] Y. R. TAUSCZIK eta J. W. PENNEBAKER, 2010, «The psychological meaning of words: LIWC and computerized text analysis methods», *Journal of language and social psychology*, **29**(1), 24–54.

[19] K. KIM, S. CHOI, J. LEE eta J. SEA, 2019, «Differences in linguistic and psychological characteristics between suicide notes and diaries», *The Journal of general psychology*, **146**(4), 391–416.

[20] B. O'DEA, M. E. LARSEN, P. J. BATTERHAM, A. L. CALEAR eta H. CHRISTENSEN, 2017, «A linguistic analysis of suicide-related Twitter posts», *Crisis*.

[21] M. J. VIOULES, B. MOULAHI, J. AZÉ eta S. BRINGAY, 2018, «Detection of suicide-related posts in Twitter data streams», *IBM Journal of Research and Development*, **62**(1), 7–1.

[22] M. M. TADESSE, H. LIN, B. XU eta L. YANG, 2019, «Detection of suicide ideation in social media forums using deep learning», *Algorithms*, **13**(1), 7.

[23] T. A. LITVINOVA, P. V. SEREDIN, O. A. LITVINOVA eta O. V. ROMANCHENKO, 2017, «Identification of suicidal tendencies of individuals based on the quantitative analysis of their internet texts», *Computación y Sistemas*, **21**(2), 243–252.

[24] A. ZIRIKLY, P. RESNIK, O. UZUNER eta K. HOLLINGSHEAD, 2019, «CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts», *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, Orrialdeak 24–33.

[25] P. L. ÚBEDA, F. M. PLAZA-DEL ARCO, M. C. DÍAZ-GALIANO, L. A. U. LOPEZ eta M. T. MARTÍN-VALDIVIA, 2019, «Detecting anorexia in Spanish tweets», *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Orrialdeak 655–663.

[26] J. A. M. MURGADO, F. M. PLAZA-DEL ARCO, P. LÓPEZ-UBEDA eta M. T. MARTIN-VALDIVIA, 2021, «A Social Monitor for Detecting Inappropriate Behavior», *Annual Conference of the Spanish Association for Natural Language Processing 2021: Projects and Demonstrations*, Orrialdeak 41–44.

[27] A. M. MÁRMOL-ROMERO, A. MORENO-MUÑOZ, F. M. PLAZA-DEL ARCO, M. D. MOLINA-GONZÁLEZ, M. T. MARTÍN-VALDIVIA, L. A. UREÑA-LÓPEZ eta A. MONTEJO-RAÉZ, 2023, «Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish», *Procesamiento del Lenguaje Natural*, **71**, 329–350.

[28] K. SIMON, 2023, «Twitter users, stats, data and trends», .
URL `https://datareportal.com/essential-twitter-stats`

[29] A. LEIS, F. RONZANO, M. A. MAYER, L. I. FURLONG eta F. SANZ, 2019, «Detecting signs of depression in tweets in spanish: behavioral and linguistic analysis», *Journal of medical Internet research*, **21**(6), e14199.

[30] A. LEIS, F. RONZANO, M. A. MAYER, L. I. FURLONG eta F. SANZ, 2020, «Evaluating behavioral and linguistic changes during drug treatment for depression using tweets in spanish: Pairwise comparison study», *Journal of Medical Internet Research*, **22**(12), e20920.

[31] D. RAMÍREZ-CIFUENTES, A. FREIRE, R. BAEZA-YATES, J. PUNTÍ, P. MEDINA-BRAVO, D. A. VELAZQUEZ, J. M. GONFAUS eta J. GONZÁLEZ, 2020, «Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis», *Journal of medical internet research*, **22**(7), e17758.

[32] C. GARCÍA-MARTÍNEZ, B. OLIVÁN-BLÁZQUEZ, J. FABRA, A. B. MARTÍNEZ-MARTÍNEZ, M. C. PÉREZ-YUS eta Y. LÓPEZ-DEL-HOYO, 2022, «Exploring the risk of suicide in real time on spanish twitter: observational study», *JMIR public health and surveillance*, **8**(5), e31800.

[33] M. STRAKA, J. HAJIC eta J. STRAKOVÁ, 2016, «UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing», *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Orrialdeak 4290–4297.

[34] D. ZEMAN, J. HAJIČ, M. POPEL, M. POTTHAST, M. STRAKA, F. GINTER, J.ÑIVRE eta S. PETROV, 2018, «CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies», *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Orrialdeak 1–21, Association for Computational Linguistics, Brussels, Belgium.
URL `https://aclanthology.org/K18-2001`

[35] J. PESTIAN, H.ÑASRALLAH, P. MATYKIEWICZ, A. BENNETT eta A. LEENAARS, 2010, «Suicide note classification using natural language processing: A content analysis», *Biomedical informatics insights*, **3**, BII–S4706.

[36] A. SARKAR, 1998, «The conflict between future tense and modality: The case of will in English», *University of Pennsylvania Working Papers in Linguistics*, **5**(2), 6.