

Gradu Amaierako Lana

Informatika Ingeniaritzako Gradua

Konputazioa

Suizidio-ideiagintzaren identifikazioa sare sozialetan

Sara Gracia Urzelai

Zuzendariak

Maite Oronoz Anchordoqui

Alicia Pérez Ramírez

2023ko ekainak 21

Laburpena

Azken urteetan suizidioa gizartearen kezka nagusietako bat bihurtu da. Gainera, sare sozialak gure egunerokotasunaren parte bilakatu dira eta emozioak adierazteko erabiltzen dira askotan. Hizkuntzaren prozesamenduaren arloan gero eta ugariagoak dira sare soziale-tako testua abiapuntutzat hartuta suizidio-ideiagintza zein beste nahasmendu psikologiko batzuen zantzuak erakusten dituzten mezu zein erabiltzaileak detektatzeko egindako saia-kerak. Proiektu honetan suizidio-ideiagintzaren ikerkuntza arloan ohikoak diren bi atazei egin zaie aurre Reddit sare sozialeko ingelesez idatzitako mezuak abiapuntutzat hartuz.

Alde batetik mezuen analisi linguistikoa gauzatu da ikasketa sakonean oinarritutako UDPipe analizatzailea erabiliz, besteak beste. Analisi honetatik ondorioztatu da suizidio-ideiagintzarekin erlazionatutako mezuak luzeagoak izan ohi direla eta lehen pertsona singularreko izenordain gehiago erabiltzen dituztela.

Bestalde, mezuen gaineko sailkapen bitarra landu da hauek suizidio-ideiagintzaren zantzuak erakusten dituzten ala ez erabakitzeko. Horretarako bi hurbilpen ezberdin eraiki dira. Alde batetik artearen egoerako sistemak erabili dira, baliabide konputazional asko kontsumitzen dituztenak. Bestetik, mezuak errepresentatzeko modu alternatibo bat propo-satu da, LDA topiko-eredua erabiltzen duena, eta simpleagoak diren sailkatzaileen gainean probatu da. Bi hurbilpen hauekin lortutako emaitzak alderatzean ondorioztatu da ataza honetan topiko-ereduak baliagarriak direla mezuak errepresentatzeko eta oinarritzkoagoa den hurbilpenarekin artearen egoerako ohiko neurona-sareen errendimendura gertura-tzea posible dela, ez ordea transformerren emaitzetara. Izan ere, transformerrek lortutako emaitzak nabarmenki gailentzen dira gainontzeko ereduak eskuratutakoetatik, % 97,9ko asmatze-tasara iritsiz (neurona-sareek lortutakoak baino 6 puntu gehiago).

Amaitzeko, egindako lanarekin ikusi da ikerketa-esparru honetan badagoela zer egina eta etorkizunerako norabide ezberdin ugari zabaldu dira, hala nola, beste ezaugarri linguis-tiko batzuk aztertzea, *emojien* erabilera ikertzea edota mezuen errepresentaziorako aldaera ezberdinak probatzea.

Gaien aurkibidea

Gaien aurkibidea	iii
Irudien aurkibidea	v
Taulen aurkibidea	vi
1 Sarrera	1
1.1 Motibazioa	1
1.2 Suizidio-ideiagintzaren definizioa	2
1.3 Proiektuaren helburuak	2
1.4 Dokumentuaren egitura	4
2 Proiektuaren plangintza	5
2.1 Lanaren Deskonposaketa Egitura	5
2.1.1 Lan-paketeak	6
2.2 Denboraren kudeaketa	7
2.2.1 Gantt diagrama	7
2.2.2 Denboren taula	7
2.3 Emangarri eta mugarriak	7
2.4 Informazio-sistema	8
2.5 Komunikazioak	8
2.6 Arriskuen analisisa	9
3 Aurrekariak	11
3.1 Antzeko atazak	11
3.2 Metodoak	12
3.2.1 Mezuen analisi linguistikoa	12
3.2.2 Mezuen sailkapena	15
3.3 Datu-sortak	16
4 Materialak eta metodoak	19
4.1 Erabilitako datu-sorta	19
4.1.1 Datu-sortaren ezaugarriak	19
4.1.2 Datu-sortarekin lan egin duten aurrekariak	21
4.2 Mezuen analisi linguistikorako metodoak	21
4.2.1 Mezuen analisi linguistikoa	22
4.2.2 Emaitzen analisisa	23
4.3 Mezuen sailkapenerako metodoak	23

4.3.1	Datuen prestaketa	23
4.3.2	Mezuen errepresentazioa	24
4.3.3	Sailkatzaileak	26
4.3.4	Ebaluazioa	28
5	Mezuen analisi linguistikoa	31
5.1	Mezuen luzera	31
5.2	Galdera-markak	32
5.3	Lehen pertsona singularreko izenordainak	33
5.4	Aditzen denborak	34
5.4.1	Lehenaldia	34
5.4.2	Orainaldia	34
5.4.3	Geroaldia	35
5.4.4	Hiru aditz-denboren arteko konparaketa	36
5.5	Analisiaren laburpena	37
6	Mezuen sailkapena	39
6.1	Esperimentuen diseinua	39
6.2	Datuen prestaketa	40
6.2.1	Aurreprozesaketa	40
6.2.2	Garbiketa	42
6.2.3	Prestaketaren eragina	42
6.2.4	Partiketa	42
6.3	Artearen egoerako hurbilpena	43
6.3.1	Errepresentazioa: hitz-embeddingak	44
6.3.2	Sailkatzaileak: ikasketa sakonean oinarrituak	44
6.3.3	Lortutako emaitzak	47
6.3.4	Eztabaida	47
6.4	Hurbilpen oinarritzkoagoa	49
6.4.1	Errepresentazioa: topiko-ereduak	49
6.4.2	Sailkatzaileak: ikasketa automatikoan oinarrituak	50
6.4.3	Lortutako emaitzak	51
6.4.4	Eztabaida	53
6.5	Sailkapenaren laburpena	54
7	Ondorioak eta etorkizuneko lana	57
7.1	Ondorio zientifiko-teknikoak	57
7.2	Proiektuaren kudeaketarekin erlazionatutako ondorioak	59
7.3	Etorkizunerako lana	60
7.4	Ondorio pertsonalak	60
A	Hurbilpen oinarritzkoagoaren esperimentazioa	63
	Bibliografia	65

Irudien aurkibidea

2.1	Proiektuaren LDE diagrama.	5
2.2	Proiektuaren Gantt diagrama.	7
3.1	LIWC softwarearen demoak eskaintako emaitzaren adibidea.	14
4.1	<i>Suicide and Depression Detection</i> datu-sortako mezuen luzera hitzetan.	20
4.2	2-dimentsioko Word2Vec errepresentazioen adibideak grafikoki.	25
5.1	Mezuen luzeraren kutxa-bibote diagrama klaseka.	32
5.2	Mezuen galdera-marka kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.	32
5.3	Mezuen lehen pertsona singularreko izenordain kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.	33
5.4	Mezuen lehenaldiko aditz kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.	34
5.5	Mezuen orainaldiko aditz kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.	35
5.6	Mezuen geroaldiko aditz kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.	36
5.7	Mezuen aditz-denboren erabileraren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.	37
6.1	Mezuen sailkapena egiteko jarraitutako esperimentazio eskema.	40
6.2	Datuen prestaketaren laburpena, aurreprozesaketa eta garbiketan jarraitutako prozesua adieraziz, pausoz pauso.	41
6.3	$m = 2$ topikoko ereduarekin lortutako topikoen hitz-hodeiak.	50
6.4	Sailkatzaile klasikoekin asmatze-tasak (%) izandako bilakaera topiko kopuruaren arabera.	51

Taulen aurkibidea

2.1	Lan-pakete bakoitzerako estimatutako denbora.	8
2.2	Proiektuaren mugarriak.	8
3.1	Suizidio-ideiagintza lantzen duten datu-sorten laburpena, sare sozialak iturri dituztenak.	16
4.1	<i>Suicide and Depression Detection</i> datu-sortaren ezaugarriak. \bar{x} : mezuen batez besteko luzera hitzetan, σ : mezuen luzeraren desbideratze estandarra hitzetan.	19
4.2	<i>Suicide and Depression Detection</i> datu-sortarekin lan egin duten aurrekariak. A.T.: asmatze-tasa.	21
4.3	2-dimentsioko Word2Vec errepresentazioen adibideak, x eta y balioak hitz bakoitzaren XY espazioko koordinatuak izanik.	25
5.1	Mezuen luzerarekin erlazionatutako estatistikoak klaseka.	32
5.2	Mezuen galdera-marka kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.	32
5.3	Mezuen lehen pertsona singularreko izenordain kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.	33
5.4	Mezuen lehenaldiko aditz kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.	34
5.5	Mezuen orainaldiko aditz kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.	35
5.6	Mezuen geroaldiko aditz kopuruarekin erlazionatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.	36
6.1	Datu-sortaren deskribapena prestaketaren fase bakoitzean.	43
6.2	Datu-sortaren partiketari buruzko datuak. OOV: <i>out-of-vocabulary</i>	43
6.3	Word2Vec hitz-embeddingekin eta artearen egoerako sailkatzaileekin lortutako emaitzak (%).	47
6.4	ELECTRA transformerrarekin lortutako emaitzen nahasmen-matrizea.	48
6.5	$m = 5, 150, 300$ topikoko LDA errepresentazioen gainean sailkatzaile klasi-koekin lortutako emaitzak (%).	52
6.6	$m = 300$ topikoko LDA errepresentazioa eta bagging teknika erabiliz eraikitako meta-sailkatzailearekin lortutako emaitzen nahasmen-matrizea.	53

7.1	LDEn definitutako adar bakoitzerako egindako estimazioak eta izandako desberapenak ordutan.	59
A.1	Topiko kopuru (m) ezberdineko LDA errepresentazioen gainean sailkatzaile klasikoekin lortutako emaitzak (%).	64

Sarrera

1.1 Motibazioa

Sare sozialetan hainbat nahasmendu psikologikorekin lotutako komunitateak daude. Horietan, milaka pertonek beren egoera emozionala partekatzen dute, laguntza eskatzen dute edo, besterik gabe, txateatu egiten dute. *Nature* aldizkarian argitaratutako ikerketa-lan batek [1] ondorioztatu zuen buruko nahasmenduak dituzten pazienteek osasun-zerbitzuei egindako premiazko laguntza-eskaerak ugari egin zirela osasun mentalarekin lotutako gaiei buruzko txio gehien argitaratu ziren egunetan.

Askotan, suizidioa buru-nahasmendu batek eragindakoa izaten da. Osasunaren Mundu Erakundearen arabera munduan urtero 700.000 pertsona baino gehiago hiltzen dira suizidioaren ondorioz. Are gehiago, 15-29 urte bitarteko gazteen arteko laugarren heriotza-kausa ohikoena izan zen 2019. urtean¹.

Duela urte gutxi, suizidioaren arrisku-faktoreen azterketaren paradigma aldatzea proposatzen zuten lanak argitaratzen hasi ziren, ikuspegi tradizionalago batetik datuen analisiaren eta ikasketa automatikoan oinarritutakora igarotzea iradokitzen zutenak. Esaterako, 2017. urtean gauzatutako meta-analisi batek [2] ondorioztatzen du azken 50 urteetan ez dela pentsamendu eta jokabide suiziden iragarpen-gaitasuna hobetu, eta argudiatzen du egindako aurkikuntzek iradokitzen dutela fokua arrisku-faktoreen azterketatik ikasketa automatikoko algoritmoetara aldatu behar dela.

Bi urte beranduago argitaratutako beste lan batean [3] antzeko hausnarketa egin zen. Bertan adierazten da beste heriotza-kausa nagusi askoren heriotza-tasek azken mendearen behean egin duten arren suizidio-tasa konstante mantendu dela. Egileen ustez suizidioaren prebentzioan dugun aurrerapenik eza, hein handi batean, arazo honen ulermen eskasaren ondorio da. Honi aurre egiteko, pentsamendu eta jokabide suizidak *in situ* aztertzekeko teknologiaren beharra azpimarratzen du. Horretarako iturri garrantzitsu bat hizkuntza da. Tradizionalki, elkarriketa elementu nagusia izan da kliniko-paziente elkarreraginean, norbanako baten osasunaren ikuspegi garrantzitsua eskaintzen baitu. Hala ere, elkarre-ragin horiek urriak eta denboran urrunak izan daitezke. Egungo testuinguruan gero eta

¹Informazio-iturria: <https://www.who.int/news-room/fact-sheets/detail/suicide>

ugariagoak dira sare sozialetan pertsonen artean gertatzen diren elkarrekintzak, suizidioaren detekzio goiztiarrean, interbentzioan eta monitorizazioan aukera berriak aurkezten dituztenak [4].

Honengatik guztiagatik, azken urteetan ugaritu egin dira hizkuntzaren prozesamendua eta ikasketa sakoneko teknikak erabiliz ideia gintza suizidaren identifikazioa ardatz duten ikerketak. Ildo horretan, *CLPsych* [5] bezalako nazioarteko erronkak antolatu dira arlo honetan aurrerapenak sustatzearren.

Proiektu honetan ikerketa-esparru honekin erlazionatutako bi ataza landu dira: alde batetik mezuen analisi linguistikoa eta bestetik mezuen sailkapena. Arlo zientifikoari dagokionez, lan hau hizkuntzaren prozesamenduko arloan kokatzen da, testuen analisia, prozesamendua, eta hauen gaineko sailkapena lantzen baitira.

1.2 Suizidio-ideia gintzaren definizioa

Suizidioaren inguruko hainbat kontzepturen definizioak nahiko lausoak dira. Ingelesez, askotan, “*suicidality*” kontzeptua (euskaratuta suizidalitate litzatekeena) erabiltzen da heriotza-kausa honekin erlazionatutako hainbat termino orokortzeko: pentsamendu suizidak, suizidio-saiakerak, suizidioa eta nork bere burua zauritzea [6]. Hala ere, 2011. urtean *US Centers for Disease Control and Prevention* agentzia estatubatuarrak argitaratutako txostenean adierazten da hitz honek barne hartzen dituen fenomenoak oso ezberdinak direla eta beraz, honen erabilera ez dela onargarria [7]. Txosten horretan, suizidioa, nork bere burua zauritzearen portaera gisa definitzen da, heriotza helburu eta ondorio duena. Suizidio-saiakera, berriz, heriotza helburu duen nork bere burua zauritzearen portaera da, ondorio gisa heriotza ez duena; eta suizidio-ideia gintzak suizidioarekin erlazionatutako pentsamendu, kontsiderazio edota plangintzak hartzen ditu barne.

Suizidio-ideia gintzarentzat unibertetsalki onartua izan den definizio zehatzik ez egoteak etengabeko erronkak eragiten ditu kliniko, ikertzaile eta hezitzaileen artean [8, 9]. Harmer et al. autoreen lanean [10] aipatzen da suizidio-ideia gintza (pentsamendu edo ideia suiziden multzo gisa ere ezagutzen dena) termino zabal bat dela, suizidioarekin eta heriotzarekin erlazionatutako kontenplazio, gurari eta kezka barne hartzen dituen fenomeno heterogeneo eta gorabeheratsua. Bertan adierazten denez, ez dago ohikotzat jo daitekeen suizidio-ideia gintza edo pentsamendurik, mota eta maila ezberdineko ideiak barne hartzen baititu: lokartzeko eta inoiz ez esnatzeko desira iheskorretatik, eldarnioek elikaturiko norberaren suntsipenarekin erlazionatutako etengabeko kezketara.

Kontuan izan behar da suizidio-ideia gintza ez dagoela suizidio-saiakerarekin era zuzenean lotuta. Estatu Batuetako 2009-2014 bitarteko datuetan oinarritutako lan batean [11] ikusi zen pentsamendu suizidak erakusten zituzten 31 pertsonetatik bakarra saiatzen zela bere buruaz beste egiten.

1.3 Proiektuaren helburuak

Proiektu honen helburu nagusia suizidio-ideia gintzak sare sozialetan duen presentzia identifikatzea eta aztertzea izan da, hizkuntzaren prozesamenduaren ikuspegitik.

Horretarako, ikerketa esparru honetako aurrekariak aztertu dira landu ohi diren atazak ezagutzeko asmoz. Bi korrante nagusi nabarmendu dira: mezuen gaineko analisi linguisti-

koan fokua jartzen dutenak eta mezuen gaineko sailkapena jorratzen dutenak. Hala ere, ez da topatu bi ataza hauek sakonki eta datu-sorta bakar baten gainean aztertzen dituen lanik. Proiektu honen bitartez hutsune hau bete nahi izan da, alor honetan egindako saiakera ezberdinak bilduz eta hauen sendotasunak probestuz eta ahuleziei aurre eginez analisi osatu bat aurkezteko.

Helburu hau bi ataza nagusitan banatu da, jarraian azaltzen den moduan:

Mezuen analisi linguistikoa egitea. Aztertutako lan ugartan ezaugarri bereziak aurkitu dira suizidio-ideiagintzaren zantzuak erakusten dituzten testu mota ezberdinetan (txioak, blogetako testuak, suizidio-oharrak...). Ezaugarri horien adibide dira mezuen luzera edota lehen pertsonako izenordainen, aditz-denbora ezberdinen edo galderamarken erabilera, esaterako. Proiektu honetan, lan ezberdinetan identifikatutako ezaugarri horiek bildu eta horietako batzuk iturritzat sare sozialak dituen datu-sorta baten gainean aztertuko dira, mezuetan bereizgarri horietako zenbat betetzen diren jakiteko.

Mezuen gaineko sailkapen bitarra egitea. Bestalde, mezuen gaineko sailkapen bitarra egiteko metodo ezberdinak landuko dira mezu batek ideiadintza suizidaren zantzuak erakusten dituen erabakitzeko. Sailkapena egitean, bi alderdi bereizten dira:

- **Mezuen errepresentazioa**

Aurrekarietan, mezuen errepresentazioaren aldetik hitz-embeddingak gailentzen dira, topiko-ereduak alde batera utziz. Lan honetan bi errepresentazio hauek erabiliko dira, bakoitzarekin lortutako emaitzak alderatzeko helburuarekin. Izan ere, proiektu honetan azaleratutako hipotesi bat ideiadintza suizida era ez-gainbegiratuan inferitutako topikoen baitan islatu daitekeela da; eta are gehiago, topikoetatik bertatik iragar daitezkeela suizidio-ideiadintzaren zantzuak.

- **Sailkatzaileak**

Sailkatzaileen kasuan, aldiz, aztertutako lanetan bi korrante nagusi bereizten dira: batetik ikasketa automatikoko sailkatzaile klasikoak erabiltzen dituztenak, eta ikasketa sakonean oinarritutako metodoak erabiltzen dituztenak, bestetik. Proiektu honetan mezuak sailkatzeko bide ezberdin hauek landuko dira datu-sorta beraren gainean, hauen emaitzak alderatu eta bakoitzaren ezaugarrien eta lortutako asmatze-tasen inguruko balorazioa egiteko.

Laburbilduz, bi **ikerketagaldere (IG)** hauek dira lan honen ardatz:

IG1. Aurrekarietako identifikatutako ezaugarri linguistikoak bereizten dira lan honetan aukeratutako datu-sortan?

IG2. Posible al da artearen egoerako hurbilpenen emaitzetara gerturatzea oinarritzkoagoa den eredu bat erabiliz?

1.4 Dokumentuaren egitura

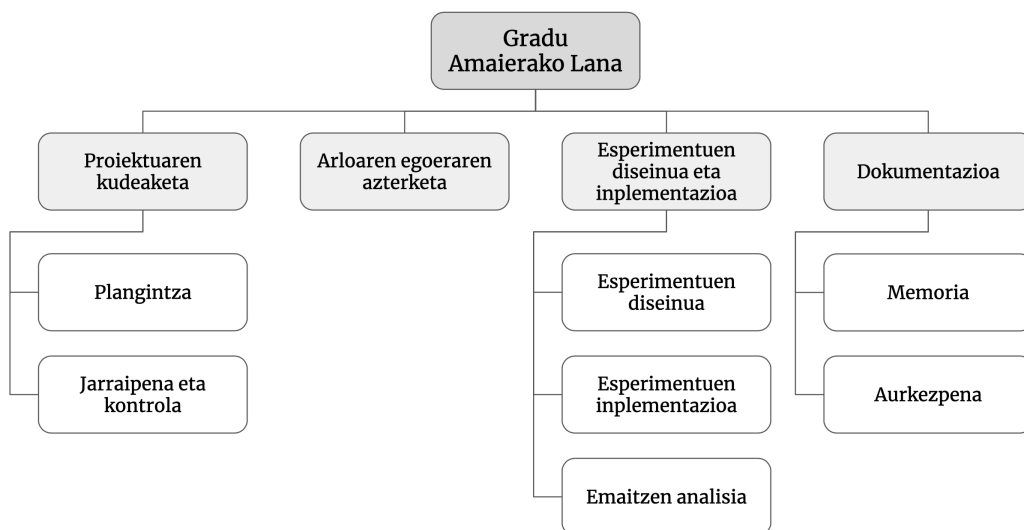
Dokumentu hau honela antolatzen da: Lehenik eta behin, 2. kapituluaren proiektuaren plangintza aurkitzen da. Jarraian, 3. kapituluaren, landutako ikerketa-arloaren artearen egoera analizatzen da. Ondoren, proiektu honetan erabilitako materialak eta metodoak aurkezten dira 4. kapituluaren. Hurrengo bi kapituluetakako bakoitzean lan honetan ebatzitako ataza bat azaltzen da: 5. kapituluaren mezuen analisi linguistikoari buruz hitz egiten da eta hurrengoan, 6. kapituluaren, mezuen sailkapenari buruz. Amaitzeko, 7. kapituluaren ondorioak aurkezten dira eta etorkizunerako proposamenak plazaratzen dira.

Proiektuaren plangintza

Gradu Amaierako Lana irismen handiko lana da, eta beraz, ezinbestekoa da plangintza on bat diseinatzea lanaren bideragarritasuna bermatzeko. Kapitulu honetan diseinatutako plangintza azaltzen da, lanaren deskonposaketa, denboraren kudeaketa, proiektuaren mugarrak, erabiliko den informazio-sistema, komunikazio motak eta arriskuen analisia barne hartzen dituelarik.

2.1 Lanaren Deskonposaketa Egitura

Gradu Amaierako Lanaren helburuak betetzen direla ziurtatzeko Lanaren Deskonposaketa Egitura (LDE) bat diseinatu da, eginkizunak lan-paketeetan banatuz. 2.1. irudian deskonposaketa honen diagrama ikus daiteke. Jarraian, adar bakoitza eta hau osatzen duten lan-paketeak azaltzen dira.



2.1 Irudia: Proiektuaren LDE diagrama.

2.1.1 Lan-paketeak

2.1.1.1 Proiektuaren kudeaketa

Adar honetan proiektuaren kudeaketarekin zerikusia duten lan-paketeak kokatzen dira, proiektuaren bizi-ziklo guztian zehar presente egongo direnak.

Plangintza. Lan-pakete honek barne hartuko ditu proiektuaren helburuak, mugarriak, deskonposaketa egitura, denboraren kudeaketa eta arriskuak zehazteko egin beharreko lanak.

Jarraipena eta kontrola. Honen bitartez proiektuaren garapena egokia dela bermatuko da, plangintzan zehaztutako helburu, denbora eta bestelakoak beteko direla ziurtatuz. Gainera, balizko desbiderapenek eragindako moldaketak egingo dira. Era berean, hemen sartuko dira proiektuaren jarraipenari buruz eztabaidatzeko egingo diren bilerak. Lan-pakete hau presente egongo da proiektuaren bizi-ziklo guztian zehar.

2.1.1.2 Arloaren egoeraren azterketa

Adar honetan ez da azpi-paketerik definitu. Lan-pakete nagusi honetan proiektuan jorratuko den ikerketa-arloari buruzko informazioa eskuratzeko beharrezko lanak kokatuko dira, hala nola, aurrekarien lanak, eskuragarri dauden datu-sortak eta atzigarri dauden proiektuak bilatzeko egin beharrekoak. Informazio hau ezinbestekoa izango da ondoren egingo diren esperimenduak zehazterako orduan.

2.1.1.3 Esperimentuen diseinua eta inplementazioa

Adar honetan proiektu honetan egingo den esperimentazioarekin lotutako lan-paketeak kokatu dira.

Esperimentuen diseinua. Esperimentazioko lehen fasea da. Ikerketa-arloaren egoera kontuan izanik eta proiektuaren helburuak aintzat hartuta, hauek betetzeko egin beharreko esperimenduak pentsatuko dira.

Esperimentuen inplementazioa. Diseinua zehaztu ostean esperimenduak inplementatuko dira.

Emaitzen analisisa. Lan-pakete honetan kokatuko da esperimentuen emaitzak analizatzeko eta horietatik ondorioak ateratzeko egin beharreko lana.

2.1.1.4 Dokumentazioa

Adar hau proiektuarekin lotutako dokumentazioari dagokio. Proiektuaren bi emangarriak, memoria eta defentsan erabiliko den aurkezpena, sortzea da honen helburua.

Memoria. Lan-pakete hau proiektuaren memoriaren idazketan oinarritzen da, emangarrietako bat izango dena. Bertan idatziko dira proiektuarekin erlazionatutako xehetasun guztiak.

Aurkezpena. Proiektuaren defentsa egiterako garaian erabiliko den euskarria izango da, lanaren sintesi argi eta ordenatua. Hau memoria osatu ostean egingo da. Gainera, aurkezpenari dagokion gidoia ere idatziko da lan-pakete honetan.

2.2 Denboraren kudeaketa

Proiektuaren plangintzan garrantzia handia du denboraren kudeaketak. Izan ere, Gradu Amaierako Lanak 12 *European Credit Transfer and Accumulation System* (ECTS) kredituko pisua du, 300 lanorduren baliokide direnak. Beraz, ezinbestekoa da alde aurretik denboraren banaketaren estimazioa egitea, baita jarraituko diren epeak finkatzea ere. Horretarako, jarraian azaltzen diren Gantt diagrama eta denboren taulak sortu dira.

2.2.1 Gantt diagrama

Gantt diagraman (2.2. irudia) lan-pakete bakoitzaren hasiera eta amaiera datak finkatzen dira. Proiektuaren kudeaketa adarra presente egongo da ia bizi-ziklo guztian, memoriaren entregara arte. Gainontzeko lan-paketeak aldiz, bata bestearen atzetik landuko dira, gutxi gorabehera. Lehenik eta behin arloaren egoeraren azterketa egingo da. Ondoren, esperimentuen diseinu eta implementazioa gauzatuko dira, modu iteratiboan, eta amaitzeko, lortutako emaitzak analizatuko dira. Memoriaren idazketa apirilean hastea estimatu da, eta aurkezpena, aldiz, memoria entregatu ostean, uztailaren hasieran, prestatuko da.

Lan-paketea		Urtarrila	Otsaila	Martxoa	Apirila	Maiatza	Ekaina	Uztaila
Proiektuaren kudeaketa	Plangintza							
	Jarraiapena eta kontrola							
Arloaren egoeraren azterketa								
Esperimentuen diseinua eta implementazioa	Esperimentuen diseinua							
	Esperimentuen implementazioa							
	Emaitzen analisisa							
Dokumentazioa	Memoria							
	Aurkezpena							

2.2 Irudia: Proiektuaren Gantt diagrama.

2.2.2 Denboren taula

Lan-pakete bakoitzerako estimatutako ordu kopurua 2.1. taulan aurki daiteke. Hau egiterako orduan kontuan izan dira alde batetik pakete bakoitzeko desbiderapen posibleak eta guztira proiektuari teoriarik eskaini behar zaion ordu kopurua, 300 ordu. Taulan letra lodiz idatzitako errenkadek Lanaren Deskonposaketa Egitura zehazterakoan definitutako adarrei dagozkie, gainean adar hori osatzen duten lan-paketeak dituztelarik.

2.3 Emangarri eta mugarriak

Proiektu honetan bi emangarri nagusi daude, memoria eta aurkezpena edo defentsa, bakoitzak entregatze- edo aurkezpen-data bat duelarik. Hala ere, hauek entregatu ahal izateko, lehenik eta behin defentsa-eskaera egin behar da. Beraz, guztira hiru mugarri izan behar dira kontuan, 2.2. taulan ikus daitezkeenak. Proiektuaren defentsaren data zehatza ez da ezagutuko defentsa eskaera onartu arte, beraz, taulan honi dagokion denbora-tartea adierazi da.

2. PROIEKTUAREN PLANGINTZA

Lan-paketea	Estimazioa (orduak)
Plangintza	10
Jarraipena eta kontrola	30
Proiektuaren kudeaketa	40
Arloaren egoeraren azterketa	70
Esperimentuen diseinua	20
Esperimentuen implementazioa	60
Emaizen analisia	15
Esperimentuen diseinua eta implementazioa	95
Memoria	80
Aurkezpena	15
Dokumentazioa	95
GUZTIRA	300

2.1 Taula: Lan-pakete bakoitzerako estimatutako denbora.

Mugarria	Epemuga
Defentsa-eskaera	2023/06/19
Memoriaren entrega	2023/06/25
Proiektuaren defentsa	2023/07/03-14

2.2 Taula: Proiektuaren mugarriak.

2.4 Informazio-sistema

Proiektu honetan erabilitako informazio-sistema bi ataletan banatzen da: implementazioari dagokiona eta memoriari dagokiona.

Implementazioa. Nahiz eta proba batzuk ordenagailuan bertan, lokalean, egin, proiektuan zehar garatutako programa guztiak IXA taldeko makinetan gordeko dira, *sgracia003* erabiltzaileari dagokion karpetan. Makina hauetan kokatzeak fitxategien segurtasun-kopiak egingo direla bermatzen du.

Memoria. Memoria Overleaf plataforman idatziko da, fitxategiak hodeian gordetzen dituen online errepositorioan. Honek bi abantaila nagusi eskaintzen ditu. Batetik, dokumentua zuzendariekin partekatzeko aukera, uneoro edozein partaidek azken bertsioa atzigarri izatea ahalbidetzen duena. Bestetik, fitxategiaren segurtasun-kopia automatikoak izatea.

2.5 Komunikazioak

Proiektuaren partaideen arteko komunikaziorako bi kanal nagusi erabiliko dira: posta elektronikoa eta bilerak.

Posta elektronikoa. Honen bitartez bilera-deialdiak eta aktak bidaliko dira beharrezkoa denean, eta kontsulta puntualak bideratzeko erabiliko da.

Bilerak. Bilerak telematikoki egingo dira, Webex plataformaren bitartez. Hauek proiektuaren jarraipenerako eta hurrengo pausoak zehazteko erabiliko dira, eta printzipioz, bi astero egingo dira, adostutako egun eta orduan.

2.6 Arriskuen analisia

Proiektu honen tamaina nahiko handia denez, honen bideragarritasuna bermatzeko ezinbestekoa da arriskuen analisia egitea eta balizko arriskuen aurrean nola jokatu pentsatzea. Honela, ustekaberen bat jazota ere, honi aurre egiteko plan bat izateak finkatutako helburu eta epeak errespetatzea ahalbidetuko luke. Jarraian, identifikatutako arrisku nagusiak azaltzen dira.

Jatorrizko plangintzarekiko desbiderapenak. Tamaina honetako proiektuetan ohikoak izaten dira plangintzarekiko desbiderapen esanguratsuak. Gerta daiteke lan-paketeren batentzat egindako denbora-estimazioa motz gelditzea eta uste baino denbora gehiago dedikatu behar izatea. Plangintza oso hertsiki diseinatua badago, honelako ezuste batek proiektuaren bideragarritasuna kolokan jarri dezake. Hau ekiditeko, paketei dagozkien denborak kalkulatzeko orduan badaezpadako tarte bat gehituko da.

Kontzeptu teorikoak ulertzeko arazoak izatea. Proiektu honetan graduan eskuratutako ezagutza erabiliko den arren, kontzeptu eta tresna berriak erabili beharko dira. Hauek barneratzerako orduan baliteke zailtasunak sortzea. Hau kontuan izango da honekin erlazionatutako paketei denborak esleitzean. Horrez gain, arazoak izan ezker biler gehigarriren bat egiteko aukera aztertuko litzateke, zuzendarien laguntzaz lanean aurrera egin ahal izateko.

Datu edo fitxategi garrantzitsuak galtzea. Erabilitako ordenagailuan arazo bat gerta daiteke, lokalean gordetako fitxategiak galtzea suposa dezakeena. Hau ekiditeko, bai memoria eta bai kode zein datuen fitxategiak gunegabean biltegiratuko dira, hodeian eta IXA taldeko makinetan.

Aurrekariak

Kapitulu honetan aurrekariak suizidio-ideiagintzaren detekzioaren inguruan egindako lana aztertzen da, sare sozialen testuinguruan eta hizkuntzaren prozesamenduaren ikuspuntutik, beti ere. Lehenik eta behin, ikerketa esparru honen barnean lantzen diren hainbat ataza aurkezten dira, proiektu honetan ebatzitakoekin antzekotasunak dituztenak. Ondoren, aurrekariak erabilitako metodoak azaltzen dira, bai mezuen ezaugarri linguistikoak aztertzeko, bai sailkapena egiteko. Horrela, proiektu honetan hauen sendotasunak aprobetxatu daitezke eta ahuleziak identifikatu hauei aurre egiteko. Amaitzeko, suizidio-ideiagintza lantzen duten hainbat datu-sorta bildu dira, iturritzat sare sozial ezberdinak dituztenak.

3.1 Antzeko atazak

Psikologian aspalditik landu izan den gaia da osasun mentaleko nahasmenduek pertsona baten hizkuntzaren erabileran duten eraginarena. Azken urteetan hizkuntzaren prozesamenduko zenbait ikerlari psikologia klinikoko erronkei aurre egiteko teknologiak garatzen ahalegindu dira [12]. Pazienteen hizkuntzaren erabilera aztertzeko tresna erabilgarria dira sare sozialak. Are gehiago, hainbat ikerketetan ikusi da erabiltzaileek sare sozialetan duten portaera eta hauek sortutako edukia beren osasun mentalaren egoeraren berri izateko erabil daitezkeela [13, 14, 15].

Horren adibide dira elikadura-nahasmenduekin erlazionatutako komunitate eta edukia identifikatzeko egindako saiakerak. Wang et al. autoreen lanean [16] Twitterreko edukia erabiltzen da erabiltzaile batek gisa horretako nahasmendu bat pairatzen duen erabakitze-ko, sailkapen bitarreko paradigma erabiliz (elikadura-nahasmendua du edo ez). Beste lan batean Instagram sare sozialaren gaineko azterketa egiten da [17]. Izan ere, Instagramek elikadura-nahasmenduak bizimodu zilegitzat jotzearen aldeko edukia ezabatzen du. Lan horretan gai horrekin erlazionatutako edukia ikertzen da, banaketa bat eginez: Instagramek ezabatutako edukia eta ezabatua izan ez dena, eta ondorioz, publiko dagoena. Ikasketa automatikoko teknikak erabiliz irudi edota bideo horien oinak aztertzen dira bereizketa horren arrazoiak aztertzeko. Chancellor et al. autoreak Instagrameko elikadura-nahasmenduekin erlazionatutako edukia oinez baliatu dira erabiltzaileen gaixotasun psikikoen larritasuna iragartzeko [15].

3. AURREKARIAK

2014. urtean sortutako *CLPsych workshopak* (www.cpsych.org) urtero nahasmendu psikologiko ezberdinak aztertzeko deialdia luzatzen dio hizkuntzaren prozesamenduko ikerlarien komunitateari. 2015. urteko edizioan proposatutako ataza depresioa edo trauma osteko estresaren nahasmendua duten Twitter sare sozialeko erabiltzaileak identifikatzea izan zen, sailkapen bitarreko hiru esperimenteren bitartez: depresioa duten erabiltzaileak versus kontrol erabiltzaileak; trauma osteko estresaren nahasmendua duten erabiltzaileak versus kontrol erabiltzaileak; depresioa duten erabiltzaileak versus trauma osteko estresaren nahasmendua duten erabiltzaileak [12].

2018. urteko edizioan helburua pertsonen uneko eta etorkizuneko osasun mentalari buruzko iragarpenak egitea izan zen, haurtzaroan egindako idazlanetatik abiatuta [18]. Horretarako, *National Child Development Study* ikerlan britainiarreko datuak erabili ziren. Ikerketa horretan pertsona talde baten osasun mentala aztertu zen jaiotzatik hasita 60 urte baino gehiagorako denbora-tartean. Beste hainbat daturen artean, parte-hartzaileek 11 urterekin egindako idazlanak bildu ziren, ostera *CLPsych* erronkan erabili zirenak.

Hurrengo urtean erronka Redditeko erabiltzaileen suizidio-arriskua iragartzea izan zen, 4 arrisku-mailako eskalan: *no risk*, *low risk*, *moderate risk* eta *severe risk* [5]. 2022an Redditeko erabiltzaileen aldarte aldaketa izan zen aztergai [19]. Horretarako hainbat erabiltzaileen mezuak erabili ziren, kronologikoki ordenatuak, helburua denbora-lerro horretan aldarte-egoera aldaketak identifikatzea izanik. Gainera, bigarren erronka gisa 2019. urteko edizioarekin uztartu eta aldarte-aldaketa horiek erabiltzaileen suizidio-arriskua iragartzeko erabiltzea proposatu zen.

Suizidio-ideiagintzaren detekzioan fokua jarrita, hainbat ataza ezberdin gauzatu izan dira. Alde batetik, mezu mailako detekzioa egiten duten lanak daude. Proiektu honetan hau lantzea erabaki da eta 3.2. atalean honetan jardun diren lanen berrikuspena egiten da. Beste ikerketa batzuetan, aldiz, erabiltzaileen portaera aztertzeko da hauen mezu-bildumez baliatuz. De Choudhury et al. autoreen lanak [20], esaterako, Reddit sare sozialeko osasun mentalarekin erlazionatutako foroetan idaztetik ideia gintza suizidaren zantzuak erakusteko trantsizioa du aztergai. Horretarako, mezuetan erabilitako hizkuntza analizatzen da metodo estatistikoak erabiliz. Gainera, erregresio logistikoaren bitartez erabiltzaile batek aipatutako trantsizioa jasango duen iragartzeko saiakera egiten da.

Badaude lan batzuk mezuen denbora-markak erabiltzen dituztenak erabiltzaileek izandako eboluzioa aztertzeko. Honen adibide da Sawhney et al. autoreen lana [21]. Bertan *STATENet* aurkezten da, denboraren kontzientzia duen transformerra. Hau, Twitterreko erabiltzaile baten txio-historia aintzat hartuz txio jakin batek suizidio-zantzuak dituen ala ez erabakitzeko erabiltzen da.

3.2 Metodoak

Atal honetan aurrekariak erabilitako metodoak eta lortutako emaitzak aurkezten dira, proiektu honetan landu diren bi atazetan fokua jarriaz: mezuen analisi linguistikoa eta mezuen sailkapena.

3.2.1 Mezuen analisi linguistikoa

Azken urteetako hainbat lanetan suizidio-zantzuak dituzten mezuen ezaugarri linguistikoak aztertu dira bereizgarriak izan daitezkeen elementuen bila. Jarraian, landu izan diren

ezaugarriak eta horretarako erabilitako tresnak azaltzen dira.

3.2.1.1 Landutako ezaugarriak

Lan askok ondorioztatu dute lehen pertsona singularreko izenordain gehiago erabiltzen direla ideia gintza suizida presente dagoen poema [22], suizidio-ohar [23], txio [24, 25] eta Redditeko mezuetan [26]. Hainbat kasutan ikusi da testu hauek heriotzarekin erlazioan-tutako hitz gehiago erabiltzen dituztela, edota honi erreferentzia gehiago egiten zaizkiola [22, 24]. Gainera, emozio negatiboekin lotutako hitzak ohikoagoak direnaren ebidentzia aurkitu da [23, 27], eta baita ezeztapena adierazteko hitz gehiago [27] eta indartzaile gehiago [25] erabiltzen direnarena ere, esandakoa biziagotzeko. Aditzei dagokienean, suizidio-oharretan geroaldiko aditz gutxiago erabiltzen direla uste da [23], eta txioetan orainaldian fokua jartzen dela [24].

Bestalde, suizidio-ideia gintzarekin erlazioan-tutako txioak luzeagoak direla [24] eta Redditeko mezuen kasuan galdera-marka gehiago erabiltzen dituztela [26] ikusi da.

Laburbilduz, aurrekariaren arabera suizidio-ideia gintzaren zantzuak dituzten iturri ezberdinetako testuek hainbat ezaugarri linguistiko bereizgarri izan ohi dituzte. Hala ere, 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 erlazioan-tutako hitz gehiago
- Emozio negatiboekin lotutako hitz gehiago
- Ezeztapena adierazteko hitz gehiago
- Indartzaile gehiago
- Geroaldiko aditz gutxiago
- Fokua orainaldian
- Mezu luzeagoak
- Galdera-marka gehiago

3.2.1.2 Erabilitako tresnak

Ezaugarri hauek aztertzeko erabilitako tresnei dagokienean, aztertutako lanen gehiengoak [22, 23, 24, 27, 28] Linguistic Inquiry and Word Count (LIWC) softwarea erabiltzen du. Tresna honek testu bateko hitz bakoitza hiztegi multzo batean bilatzen du, honi psikologiarekin lotura duen kategoria bat esleitzeko. Emaiza gisa, kategoria bakoitzarekin lotutako testuko hitz guztien portzentaia itzultzen du [22].

Softwareak eskaintzen dituen emaitzak hobe ulertzeko asmoz proba bat egin da jarraian aurkezten den mezuarekin:

3. AURREKARIAK

Life isn't doing it for me anymore I feel so lost in the world. I wish it was all over. I wish things would just go away. I hate feeling like this. I want help and trust me i've tried. So many meds doctors and more meds. It's been three years. Still nothing the only thing i'm left with is a prescription stimulant problem. Which fucking sucks. It eats me inside and i have such a high tolerance. I don't know life just doesn't do it for me anymore. The highs aren't enough. I want out but i don't want to tell my parents. I'd rather them have to deal with my death then see there faces if i came out about any substance problems. My life revolves around getting high/drunk. I hate this i hate this so much. If i don't have something to live for in the next 10 or so days i'm putting a gun to my head. I'll find a way.

Mezu hau LIWC tresnaren demoarekin analizatuta 3.1. irudiko emaitza lortu da. Bertan, mezuak hainbat kategoriatan (ni-arekin erlazionatutako hitzak, tonu positibo eta negatiboa, hitz sozialak, prozesu kognitiboak, erakargarritasuna eta moralizazioa) lortu duen puntuazioa ikus daiteke (“*Your Text*” zutabea), baita sare sozialetako testuek lortu ohi duten puntuazioa ere (“*Average for Social Media Language*” zutabea).

Traditional LIWC Dimension	Your Text	Average for Social Media Language
I-words (I, me, my)	14.12	5.44
Positive Tone	0.56	5.93
Negative Tone	3.39	2.34
Social Words	1.13	6.74
Cognitive Processes	11.30	8.86
Allure	9.60	8.62
Moralization	0.00	0.27

3.1 Irudia: LIWC softwarearen demoak eskaintako emaitzaren adibidea.

LIWC hiztegietan oinarritutako tresna denez, ez du testuingurua kontuan hartzen eta ondorioz, hainbat kasutan akatsak egin ditzake. Software honen webgunean adierazten da honen adibide bat, “*mad*” hitzari dagokiona. Hitz hau haserrearen hiztegian kokatzen da LIWCn, baina batzuetan gozamina edo eromena adierazten du, zentzu positibo batean: “*he's mad for her*” (beregatik zoratuta dago).

Artearen egoeran aipatutako beste lan batek, berriz, aztertutako klase bakoitzeko (suizidio-ideiagintzaren zantzuak dituzten mezuak eta ez dituztenak) 200 unigrama eta bigrama ohikoenak aztertzen ditu bestelako tresnarik erabili gabe hizkuntzan ezberdintasunak aurkitzeko asmoz [26].

Esparru honetan, ordea, hutsune nabarmen bat aurkitu da. Izan ere, ikusi da ez dela ohikoa mezuen analisi linguistikoa egiteko ikasketa sakonean oinarritutako analizatzaileak erabiltzea.

3.2.2 Mezuen sailkapena

Ideiagintza suizidaren eta sare sozialen arloan oso ohikoa den ataza da sailkapena. Hau, funtsean, ikasketa automatiko zein sakoneko algoritmoen bitartez suizidio-ideiagintzaren zantzuak erakusten dituzten mezu edo erabiltzaileak detektatzeko helburuarekin egiten da.

3.2.2.1 Errepresentazio numerikoa

Ikasketa automatiko edota sakonaren bitartez testua aztertu ahal izateko beharrezkoa da hauen errepresentazio numerikoa lortzea. 2022. urtean argitaratutako lan batean [29], sare sozialak eta suizidioaren gaia ikasketa automatikoaren ikuspegitik jorratzen duten lan asko bildu eta azaltzen dira. Berrikuspen bibliografiko horretan aztertutako lan gehienetan Word2Vec [30] gisako hitz-embeddingak, hitzen zenbakizko errepresentazio sakonak, erabiltzen dira mezuak numerikoki errepresentatzeko, edota n -gramak, ondoz ondoko n hitzen sekuentziak direnak. Beste lan batzuek hitzen agerpen-maiztasunekin zerikusia duten metrikak, Term Frequency – Inverse Document Frequency (TF-IDF) edota LIWC softwareak eskainitakoak, esaterako, erabiltzen dituzte.

Helburu honekin topiko-ereduak erabiltzen dituzten lanak, aldiz, gutxi dira (eredu hauei buruzko informazioa 4.3.2.2. atalean aurki daiteke). Batzuetan ezaugarri iragarle bezala erabiltzen dira beste hainbat elementurekin batera (estatistikoak, hitz-embeddingak, TF-IDF metrikaren emaitzak, LIWC softwareak eskainitako informazioa...) [31, 32]. Proiektu honetan, besteak beste, teknika honek mezuen errepresentaziorako izan dezakeen ahalmena aztertu nahi izan da.

Jashinsky et al. autoreen lanean [33] txioetan erabiltzen diren hitzak *National Institute of Mental Health* erakundeak eta Lewinsohn et al. autoreek [34] identifikatutako suizidioaren 12 arrisku-faktorerekin lotzeko saiakera egiten da. Beste lan batzuk ere saiatu dira Redditeko mezuetan zein txioetan agertzen diren hitzak multzokatze-algoritmoak edota topiko-ereduak erabiliz multzokatzen eta ondoren 12 arrisku-faktore horiekin erlazioatzen [35, 36]. Horietako batean [36] topiko-eredu ezberdinak erabiltzen dira txioak errepresentatzeko. Horrela, autoreak gai izan dira aztertutako 12 arrisku-faktoreetatik 7 lortutako topikoetan identifikatzeko. Lan horretan ebatzi nahi den ataza Twitterreko erabiltzaile baten suizidio-arriskua altua edo baxuagoa den erabakitzea da. Errepresentazio horretatik sailkapen-zuhaitza inferituz % 84,8ko doitasuna lortu dute egileek. K-Means [37] multzokatze-algoritmo ez-gainbegiratuarekin, aldiz, autoreek ez dute ataza era onargarrian ebatzea lortu.

3.2.2.2 Sailkatzaileak

Sailkatzaileei dagokienean, bi multzo handitan sailka daitezke: Sailkatzaile klasikoak (ikasketa automatikoan oinarrituak) eta ikasketa sakonean oinarritutakoak.

Ikasketa automatikoan oinarritutako sailkatzaileen artean ataza honetarako erabilienak izan direnak Support Vector Machine (SVM), Random Forest eta erregresio logistikoa dira [29]. Rabani et al. autoreen lanean [38], esaterako, ikasketa automatikoko metodo ezberdinak (Multinomial Naïve Bayes, J48, erregresio logistikoa, SMO, Random Forest, etab.) erabiltzen dira txioak suizidioarekin erlazioa duten ala ez estimatzeko.

Ikasketa sakonean oinarritutako sailkatzaileen artean, aldiz, esparru honetan Long Short Term Memory (LSTM) eta Convolutional Neural Network (CNN) neurona-sareak dira

erabilienak [29]. Azken urteetako lan batzuek transformerrak erabiltzen dituzte. Horietako batean [39], adibidez, Google enpresak sortutako Bidirectional Encoder Representations from Transformers (BERT) sailkatzailearen aldaera ezberdinak erabiltzen dira txio batean ideia-gintza suizidarik dagoen ala ez erabakitzeko, % 95,4ko asmatze-tasa lortuz.

Erabilitako ikasketa paradigmatari dagokionean, Abdulsalam and Alhothali autoreek egindako berrikuspenean [29] aztertutako lan gehienek eredu gainbegiratuak erabiltzen dituzte mezuen sailkapenerako. Ikasketa gainbegiratuan datu etiketatutako erabiltzen dira datu eta etiketen arteko erlazioak ikasteko. Ikasketa ez-gainbegiratuan, aldiz, etiketatu gabeko datuak erabiltzen dira. Era honetako metodoek datuen egiturak aztertzen dituzte hauek erakusten dituzten antzekotasunak identifikatzeko. Hauek ere sailkapena egiteko molda daitezkeen arren, lan honetan fokua ikasketa gainbegiratuan oinarritutako sailkatzaileetan jartzea erabaki da, emaitza onenak eman dituztenak erabiliz.

3.3 Datu-sortak

Aurrekarien analisisan aztertutako lan gehienetan, autoreek lanetarako espresuki egindako bilketetatik eskuratutako datuak erabiltzen dira. Bilketa horiek sare sozialetan hitz-gakoak erabiliz [25, 33, 36] edota suizidioarekin erlazionatutako foroetako mezuak bilduz [35] egin ohi dira. Datu-sorta horiek, ordea, ez dituzte gerora argitaratzen. Honen arrazoa datu-babesa eta erabiltzaileen pribatutasuna babestu beharra izan daiteke.

Hala ere, datu-bilketa bat egitea proiektu honen irismenetik kanpo gelditzen denez, Kaggle, GitHub eta antzerako errepositorioetan bilaketa egin da suizidio-ideia-gintzari buruzko corpus publikoak aurkitzeko asmoz. Bilaketa horren emaitza 3.1. taulan laburbildu da. Orokorrean, aurkitutako datu-sortetan Twitter edota Reddit sare sozialetako mezuak biltzen dira, eta ia denetan ingelesez idatzitako mezuak soilik aurki daitezke. Ez da bakarrik gazteleraz idatzitako mezuen datu-sorta publikorik topatu, ezta euskarazkorik ere.

Datu-sorta	Iturria	Hizkuntza	Tuplak
Life! Corpus	S. sozialak	Ing + Gaz	102
Detection of Suicide Ideation in Twitter using Machine Learning	Twitter	Ingelesa	9.206
Identifying Depression and Suicide displays on Twitter	Twitter	Ingelesa	999
AI For Social Good	Reddit + Twitter	Ingelesa	11.339
Suicide and Depression Detection	Reddit	Ingelesa	232.074
Reddit C-SSRS Suicide Dataset	Reddit	Ingelesa	500
The University of Maryland Reddit Suicidality Dataset	Reddit	Ingelesa	22.258

3.1 Taula: Suizidio-ideia-gintza lantzen duten datu-sorten laburpena, sare sozialak iturri dituztenak.

Life! Corpus datu-sorta sare sozial ezberdinetatik eskuratutako ingelesez idatzitako 71 mezuz eta gaztelerazko 31 mezuz osatuta dago. Mezu hauek eskuz etiketatutako izan dira, eta bi irizpide ezberdinen arabera sailkatu dira. Alde batetik, arrisku-maila bat ezarri zaie 4 graduko eskala bat erabiliz: arriskurik ez (*no risk*), premiazkoa (*urgent*), posiblea (*possible*) edo berehalakoa (*immediate*). Bestetik, mezuaren gaia edo tematika adierazi da:

suizidio-ideiagintza, tristura, depresioa, ironia... Mezu horietako batzuek erabiltzaileei buruzko informazioa biltzen dute, ez denek.

Detection of Suicide Ideation in Twitter using Machine Learning datu-sortan 9.206 mezu biltzen dira, etiketa gisa 1 dutenak mezuak suizidio-ideiagintzaren zantzuak erakusten dituzten eta 0 kontrako kasuan. Ez dago mezu horiek idatzi dituzten erabiltzaileei buruzko informaziorik. Etiketatzeko-prozesuari dagokionez, ez da informaziorik aurkitu, beraz, ezin daiteke jakin mezuak automatikoki etiketatuta diren irizpideren bati jarraituz ala aditu baten lana izan den.

Identifying Depression and Suicide displays on Twitter sorta sortzeko, autoreek Twitter sare sozialeko mezuak bildu dituzte non “*depression*” edota “*depressed*” hitzak agertzen diren. Beraz, nahiz eta datu-sortaren izenburuan suizidioarekin lotutako mezuak biltzen direla aipatu, benetan datu-sortaren gaia depresioa dela esan daiteke. Ez da aurkitu mezu horien etiketatzeari buruzko informaziorik. Mezu bakoitzeko hainbat informazio gordetzen da: erabiltzailearen izena, Twitter kontuaren izena eta deskribapena, lagun eta jarraitzaile kopurua, kokapena, mezua eta etiketa (0: mezuak ez du erabiltzaileak depresioa duenaren zantzurik erakusten, 1: mezuak erabiltzaileak depresioa izan dezakeela erakusten du). Nahiz eta erabiltzaileari buruzko informazioa gorde, ez dira erabiltzaile beraren hainbat mezu gordetzen eta beraz, datu-sorta hau ezingo litzateke erabiltzaileen sailkapena egiteko erabili.

AI For Social Good proiektuan Reddit eta Twitterreko mezuak erabiltzen dira. Zehazki, suizidio-ideiagintzaren zantzuak erakusten dituzten Redditeko 2.958 mezu aurki daitezke bertan, “*SuicideWatch*”, “*depression*” eta “*Anxiety*” *subreddit* edo foroetatik eskuratua; suizidioarekin zerikusia duten 3.000 txio, “*suicide*”, “*die*” eta “*end my life*” hitz-gakoak erabiliz bildutakoak; eta suizidioarekin erlaziorik erakusten ez duten Redditeko 5.381 mezu, “*movies*”, “*popular*”, “*books*” edota “*Jokes*” foroetan idatziak. Mezuak automatikoki etiketatuta dira, datu-bilketarako erabilitako irizpidearekin. Mezu hauek ez dute erabiltzaileei buruzko informaziorik.

Suicide and Depression Detection datu-sortan Reddit sare sozialetik eskuratutako 232.074 mezu biltzen dira. Mezu hauek “*SuicideWatch*” eta “*teenagers*” foroetatik eskuratu dira. Etiketatzeko automatikoa egin da, jatorritzat lehenengo foroa duten mezuei *suicide* etiketa esleitu zaie, eta *non-suicide* bigarren forotik lortutakoei. Klase-banaketari erreparatu, mezuak orekatuta daude. Datu-sorta honetan ere ez da erabiltzaileei buruzko informaziorik gorde.

Reddit C-SSRS Suicide Dataset datu-sortan Reddit sare sozialeko 500 erabiltzailearen mezuak biltzen dira. Erabiltzaile bakoitzeko hainbat mezu gordetzen dira eta 5 etiketa hauetako bat esleitzen zaio erabiltzaileari: *supportive*, *attempt*, *behavior*, *ideation* edo *indicator*. Etiketa hauek lau adituk esleitu dituzte, % 73ko talde-akordioa lortuz [40]. Nahiz eta mezuen data ezagutzen ez denez erabiltzaileen denbora-lerroa sortzeko aukerarik ez izan, datu-sorta honen bitartez posible litzateke erabiltzaileen gaineko sailkapena egitea.

The University of Maryland Reddit Suicidality Dataset 2019. urteko *CLPsych* nazioarteko erronka [5] gauzatzeko erabili zen datu-sorta da. Erronkaren helburua suizidio-arriskua erakusten zuten erabiltzaileak identifikatzea zen. Datu-sortan Reddit sare sozialeko 22.258 erabiltzailearen mezuak biltzen dira. Horietako 11.129k “*SuicideWatch*” foroan idatzi dute, eta gainontzeko 11.129k berriz, ez. Gainera, datu-sorta horretatik eratorritako bi azpimultzo daude eskuragarri. Batean 1.242 erabiltzaile biltzen dira (621 “*SuicideWatch*” foroan idatzi dutenak eta 621 bertan idatzi ez dutenak), zeintzuk *crowdsourcing* erabiliz etiketatuta diren.

3. AURREKARIAK

Bestean adituek etiketatutako 490 erabiltzailearen datuak gordetzen dira (245 “*SuicideWatch*” foroan idatzi dutenak eta 245 bertan idatzi ez dutenak). Bi kasuetan 4 arrisku-maila hauetako bat esleitu zaie etiketa gisa: *no risk*, *low risk*, *moderate risk* edo *severe risk*. Mezu bakoitza argitaratutako uneari buruzko informazioa gordetzen denez, posible litzateke erabiltzaileen denbora-lerroa eraikitzea. Datu-sorta hau ez da publikoa, beharrezkoa da eskaera bat egitea hau eskuratzeko.

Materialak eta metodoak

Aurrekariak suizidio-ideiagintza eta sare sozialen ikerketa-esparruan egindako lana analizatu ostean, kapitulu honetan proiektu honetan erabilitako datu-sorta aurkezten da, baita mezuen analisi linguistikoa eta sailkapena egiteko erabilitako ereduak eta ebaluaziorako tresnak ere.

4.1 Erabilitako datu-sorta

Lan honetan *Suicide and Depression Detection*¹ datu-sorta erabili da. Honen arrazoia, aztertutako datu-sorta guztietatik mezu kopuru handienekoa dela izan da. Izan ere, datu-sorta geroz eta aberatsagoa izan, orduan eta aukera gehiago zabaltzen dira bai sailkapenerako eredu handiak entrenatzeko, bai mezuen analisi linguistikoan lortutako emaitzak orokortzeko.

4.1.1 Datu-sortaren ezaugarriak

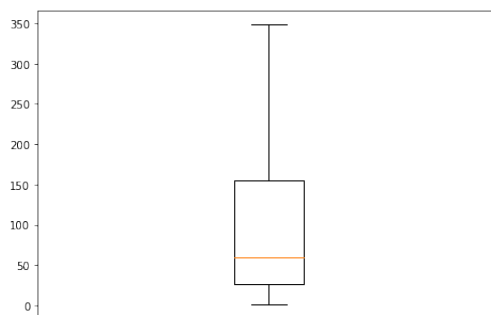
Datu-sorta hau Reddit (www.reddit.com) sare sozialeko “*SuicideWatch*” eta “*teenagers*” *subreddit* edo foroetatik eskuratutako mezu-bilduma da. Guztira, ingelesez idatzitako 232.704 mezu biltzen dira bertan. Mezu bakoitzak etiketa bat du esleituta: *suicide*, mezuak suizidio-ideiagintzaren zantzuak dituenean edo *non-suicide*, mezua gai horrekin erlazionatuta ez dagoenean. Datu-sortaren hainbat ezaugarri 4.1. taulan bildu dira: mezu kopurua, klase banaketa, mezuen batez besteko luzera hitzetan (\bar{x}), luzeraren desbideratze estandarra hitzetan (σ) eta hiztegiaren tamaina.

Mezu kop.	Klase banaketa	Luzera (\bar{x})	Luzera (σ)	Hiztegiaren tam.
232.704	<i>suicide</i> : % 50 (116.352) <i>non-suicide</i> : % 50 (116.352)	132	217	521.930

4.1 Taula: *Suicide and Depression Detection* datu-sortaren ezaugarriak. \bar{x} : mezuen batez besteko luzera hitzetan, σ : mezuen luzeraren desbideratze estandarra hitzetan.

¹Datu-sorta eskuragarri hemen: <https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch>

Taulan ikus daitekeenez mezuen luzerak desbideratze oso handia erakusten du. Honi buruzko informazio zehatzagoa eskaintzearen luzerari dagokion kutxa-bibote diagrama eraiki da, 4.1. irudian ikusgai dagoena.



4.1 Irudia: *Suicide and Depression Detection* datu-sortako mezuen luzera hitzetan.

Datu-sorta hau automatikoki etiketatua izan da, hau da, etiketak ez ditu aditu batek eskuz esleitu. “*SuicideWatch*” foroaren helburua pentsamendu suizidak dituen orok laguntza eskatu ahal izateko espazio bat izatea da. Bertan adierazten denez, foroan eskaintzen dena “berdinen laguntza pentsamendu suiziden aurka borrokatzen den edonorentzat” da. Gainera, moderatzaileak daude bertan, foroan argitaratzen den edukia kontrolatzeaz arduratzen direnak. Horregatik, printzipioz, bertako mezu guztiak suizidio-ideiagintzarekin lotuta egotea espero da. Beraz, foro honetatik eskuratutako mezuei *suicide* etiketa esleitu zaie. “*teenagers*” foroa, aldiz, nerabeei zuzendutako espazioa da, “oro har nerabeen interesekoak izan daitezkeen gai eta memeei buruz eztabaidatzeko” gunea. Kasu honetan, bertako mezuei *non-suicide* etiketa esleitu zaie, baina suizidioa nerabeen artean ere oso presente dagoenez, baliteke bertako mezuren bat gai honekin erlazionatuta egotea. Jarraian, datu-sortako bi mezu aurkezten dira adibide gisa, etiketa ezberdina dutelarik:

suicide: *Life isn't doing it for me anymore I feel so lost in the world. I wish it was all over. I wish things would just go away. I hate feeling like this. I want help and trust me i've tried. So many meds doctors and more meds. It's been three years. Still nothing the only thing i'm left with is a prescription stimulant problem. Which fucking sucks. It eats me inside and i have such a high tolerance. I don't know life just doesn't do it for me anymore. The highs aren't enough. I want out but i don't want to tell my parents. I'd rather them have to deal with my death then see there faces if i came out about any substance problems. My life revolves around getting high/drunk. I hate this i hate this so much. If i don't have something to live for in the next 10 or so days i'm putting a gun to my head. I'll find a way.*

non-suicide: *I'm not a prophet but... I know the person reading this is breathing Happy new years.*

Datu-sortaren urre-patroia fidagarria den ikusteko honen zarata-maila neurtzea erabaki da. Horretarako, 200 mezu aukeratu dira ausaz (100 klase bakoitzeko) eta gaiari buruzko jakintza askorik izan gabe, balizko etiketak esleitu zaizkie ondoren benetakoekin alderatzeko. Lortutako adostasun-maila % 96,5koa izan da, beraz, ondoriozta daiteke sailkapenerako

erabiliko diren sistemek ez dutela zarata handirik ikasiko. Adibide gisa, jarraian urrepatroiaren arabera suizidio-zantzuak erakusten dituen baina desadostasuna sortu duen mezu bat aurkezten da:

went on twitch and talked bad to a streamer i went on twith and called a girl all sorts of things she got really mad and her chat against me. i'm truly alone

4.1.2 Datu-sortarekin lan egin duten aurrekariak

Horretaz gain, datu-sorta hau erabiltzen duten hainbat lan aztertu dira, 4.2. taulan laburbildu direnak. Esan beharra dago ez dela datu-sorta honekin lan egiten duen artikuluzientifikorik aurkitu eta taulan adierazitakoak Kaggle (www.kaggle.com) eta Hugging Face (www.huggingface.co) plataformetan aurkitutako proiektuei dagozkiela. Lan bakoitzerako, taulan, honen izenburu estekatua, analisi linguistikoa landu den, eta sailkapenerako erabilitako errepresentazioa eta asmatze-tasarik altuena lortu duen sailkatzailea eta asmatze-tasa hau adierazi dira.

Lana	Analisi ling.	Errepresen.	Sailkatzailea	A.T.
Suicide Prevention	Ez	Doc2Vec	Erreg. logistikoa	% 92,2
Predicting Suicide and Word Analysis	Bai	Bag of words	Random Forest	% 83,7
Suicide Detection Using Reddit Data	Ez	TF-IDF	Meta-sailkatzailea	% 93,2
Suicidal-BERT	Bai	ELECTRA	ELECTRA	% 97,9

4.2 Taula: *Suicide and Depression Detection* datu-sortarekin lan egin duten aurrekariak. A.T.: asmatze-tasa.

Aztertutako lanetan analisi linguistikoa ia alde batera utzi da. *Predicting Suicide and Word Analysis* lanean hitzek *suicide* etiketarekin duten korrelazioa aztertu da, eta *Suicidal-BERT* lanean mezuen polaritatea eta klase bakoitzeko 100 hitz eta 20 bigrama usuenak ikusita suizidio-ideiagintzarekin erlazionatutako mezuek emozio negatibo gehiago aipatzen dituztela ondorioztatu da. Bigarren honetan *suicide* etiketadun mezuek luzeagoak direla ere ikusi da. Orokorrean, esan daiteke orain arte ez dela datu-sorta honen analisi linguistikoa sakondu, eta beraz, proiektu hau litzatekeela horretan lehena.

Sailkapenean, aldiz, emaitza esanguratsuak lortu dira, asmatze-tasarik altuena % 97,9koa izanik, ELECTRA transformerrarekin eta eredu honek erabiltzen duen errepresentazioarekin lortua *Suicidal-BERT* lanean. Hala ere, emaitzak ez dira guztiz konparagarriak, ebaluazio-eskema gisa denek *hold-out* erabili badute ere partiketa ezberdinak egin dituztelako eta *Predicting Suicide and Word Analysis* eta *Suicide Detection Using Reddit Data* lanetan datu-sortaren azpimultzo bat soilik landu delako, kasu bakoitzean ezberdina izan dena.

4.2 Mezuen analisi linguistikorako metodoak

Atal honetan mezuen analisi linguistikoa gauzatzeko jarraitutako metodoak eta prozesua azaltzen dira. Prozesu hori bi pausotan banatzen da:

1. Mezuak analizatu hauetatik datu kuantitatiboak lortzeko.

2. Datuak metodo estatistikoak erabiliz laburbildu eta emaitzak interpretatu.

Bi pauso hauetan tresna ezberdinak erabili dira, jarraian aurkezten direnak.

4.2.1 Mezuen analisi linguistikoa

Lan honetan era ezberdinetako ezaugarri linguistikoak (puntuazio-markak, mezuen luzera, kategoria gramatikalak eta aditz-denborak) analizatzea erabaki denez, tresna ezberdinak erabili behar izan dira, jarraian aurkezten direnak.

4.2.1.1 `split()` eta `len()` funtzioak

Mezuen luzera (mezuen hitz kopurua) aztertzeko, Python programazio-lengoaiak eskaintzen duen `split()` funtzioa erabili da. Honek, karaktere-kate bat zatitzeko balio du, zatitzeko irizpide bezala zuriunea hartuz. Honela, mezuak hitzetan banatzea lortu da, hitz horiekin zerrenda bat osatuz. Mezuen luzera ezagutzeko hitz-zerrenda horien luzera lortu da `len()` funtzioaren bitartez.

4.2.1.2 Bateragarritasun zehatza

Bateragarritasun zehatzak (*exact-match*) elementu jakinen agerpenak aurkitzean datza. Kasu honetan, hau erabili da mezuen galdera-marka kopurua zein den ezagutzeko.

4.2.1.3 UDPipe

Izenordainak eta aditz-denborak aztertzeko, aldiz, tresna aurreratuagoak behar dira, analizatzaile linguistikoak edo hiztegiak, esaterako.

Aurrekarien analisia egitean azaldu bezala, aztertutako lan gehienek ezaugarri hauek analizatzeko Linguistic Inquiry and Word Count (LIWC) softwarea erabiltzen dute (ikus [3.2.1.2](#). atala). Hala ere, hiztegiaren oinarritutako tresna honek dituen ahuleziak kontuan izanda eta proiektu hau hizkuntzaren prozesamenduaren testuinguruan kokatua egonik, LIWC alde batera utzi eta ikasketa sakonean oinarritutako analizatzaileen alde egitea erabaki da, hitzen kategoria gramatikala eta ezaugarri linguistikoak jaso eta hizkuntzaren aldetik analisi sakonagoa egiteko aukera eskaintzen dutelako.

Zehazki, UDPipe erabili da izenordainak eta aditz-denborak analizatzeko. 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 [41]. Gainera, *Universal Dependencies* (UD) proiektuan erakitako *treebank* gehienetarako ereduak daude eskuragarri. Kasu honetan, ingeleserako prestatutako eredu bat erabili da, *English Web Treebank* corpusa oinarri hartuta entrenatua.

UDPipe nazioarteko *Multilingual Parsing from Raw Text to Universal Dependencies* erronkaren testuinguruan sortu zen, CoNLL konferentziak antolatua [42]. Honengatik, ereduak 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 menpekoea den POS (XPOS) etiketak, ezaugarri morfologikoak edota hitzen arteko dependentzia-erlazioak.

Lan honetan, zehazki, lehendabizi mezuak tokenizatu dira eta ondoren, izenordainak eta aditz-denborak aztertzeko, token horien lemak, XPOS etiketak eta dependentzia-erlazioak izan dira kontuan.

4.2.2 Emaizten analisia

Emaizak analizatu ahal izateko beharrezkoa da mezuen azterketan lortutako datuak laburbiltzea. Horretarako, hainbat estatistiko kalkulatu dira: batezbestekoa (\bar{x}), desbiderapen estandarra (σ), kuartilak (% 25, % 50 eta % 75), eta minimoa (min) eta maximoa (max). Estatistiko hauek maiztasun absolutuen eta mezuen luzerarekiko [0-100] eskalan normalizatutako datuen gainean kalkulatu dira, bi klaseen arteko konparaketa ahalik eta justuena egiteko. Neurri hauek guztiak taula batean bildu dira ezaugarri bakoitzeko. Taulaz gain, luzerarekiko normalizatutako balioak oinarri hartuta kutxa-bibote diagrama eraiki da Pythoneko `matplotlib` liburutegia erabiliz, emaitzak era bisualago batean erakusteko asmoz.

4.3 Mezuen sailkapenerako metodoak

Atal honetan, proiektu honetan ebatziko den bigarren atazari aurre egiteko, mezuak sailkatzeko, erabilitako metodoak azaltzen dira. Sailkapenerako erabilitako metodologia 4 pausotan banatzen da:

1. Datuak prestatu.
2. Mezuen errepresentazioa lortu.
3. Sailkatzaileak entrenatu.
4. Ereduak ebaluatu.

Jarraian pauso horietako bakoitza azaltzen da, baita bakoitzean erabilitako tresnak ere.

4.3.1 Datuen prestaketa

Sailkapena egin baino lehen, hainbat arrazoiengatik, datuak prestatu dira. Alde batetik, sare sozialak testuinguru informaltzat hartzen direlako eta ondorioz esparru honetan aurkitzen den testua, normalean, desegituratuagoa izaten delako beste iturri batzuetatik eskuratutakoa baino. Bestalde, datu-sortaren tamaina nahiko esanguratsua denez, ereduaren entrenamendu prozesuaren zama konputazionala arintzeko hiztegiaren tamaina murriztearen eta mezuaren egon daitekeen zarata ezabatzearen alde egitea erabaki da.

Prozesu honetan, Tausczik and Pennebaker autoreen irizpidea erabili da [22]. Egile hauek diote hitzak bi multzo handitan sailka daitezkeela beren propietate psikometriko eta psikologikoen arabera: eduki-hitzak (*content words*) eta estilo- edo funtzio-hitzak (*style or function words*). Lehen kategoriakoak dira izenak, aditz erregularrak eta hainbat adjektibo eta aditzondo. Bigarrenekoak, berriz, izenordainak, preposizioak, artikulua, lokailuak, aditz laguntzaileak eta tankerakoak izango lirake. Hiztegi ingelesak 100.000 hitz inguru dituen arren, funtzio-hitzek honen % 0,05 osatzen dute soilik (500 hitz inguru). Hala ere, hitz hauek ahozko zein idatzizko komunikazioaren % 55 inguru suposatzen dute. Perspektiba

psikologikotik, funtzio-hitzez pertsonak nola komunikatzen diren islatzen dute, eta eduki-hitzez, aldiz, zer komunikatu nahi den adierazten dute. Ataza honetan bigarren hauetan zentratu nahi izan da, hau da, mezuen edukian jarri da fokua. Zerri buruz hitz egiten dute suizidio-ideiagintza erakusten duten mezuak? Horren araberrako sailkapena egitea posible al da?

Datuen prestaketarako erabilitako prozesua *Suicidal-BERT*² proiektuko autoreek erabilitakoaren erreproduktzioa da. Bertan Python lengoaiarako eskuragarri dauden hainbat liburutegi erabili dira, horien artean: SymSpell akats ortografikoak zuzentzeko, unidecode azentuak ezabatzeko, contractions kontrakzioak zabaltzeko eta spaCy hitzen tokenizazio eta lematizaziorako.

4.3.2 Mezuen errepresentazioa

Proiektu honetan landu diren testu-sailkapeneko metodoak erabiltzeko ezinbestekoa da lehenik eta behin mezuak numerikoki errepresentatzea. Kasu honetan bi errepresentazio modu ezberdin landu dira, jarraian azaltzen direnak.

4.3.2.1 Hitz-embeddingak

Ikerketa arlo honetan, 3.2.2.1. atalean azaldu bezala, hitz-embeddingak dira mezuak errepresentatzeko erabilienak. Arrazoi honengatik lan honetan erabili den hurbilpenetako bat izan da.

Hitz-embeddingak hitzen zenbakizko errepresentazio sakonak dira, hau da, hitzak zenbakizko bektore bihurtzen dituzte, ondoren neurona-sareen sarrera gisa balio dutenak. Bektore hauen bitartez, ereduak hitzen inguruko informazio semantiko zein sintaktikoa bereganatzen dute. Lan honetan, zehazki, Word2Vec ereduak [30] erabiliz lortu dira bektore horiek, erabiliena baita sare sozialetako suizidio-ideiagintzaren detekzioaren esparruan [29]. Eredu honen eraginkortasuna hitz baten antzekoen bektoreak elkartzean eta honen esanahiari buruzko estimazioak egitean datza, testuan dituen agerpenetan oinarrituta. Estimazio hauen bitartez hitzen arteko erlazioak identifika daitezke [43].

Demagun 2-dimentsioko Word2Vec errepresentazio bektoreak (x eta y koordinatuenak) lortu ditugula “ardoa”, “garagardoa”, “katua” eta “txakurra” hitzentzat, 4.3. taulan ikus daitezkeenak. Bektore hauek 2-dimentsioko XY espazioan irudikatu dira, 4.2. irudiko emaitza lortuz. Bertan ikus daitekeenez, “katua” eta “txakurra” hitzak elkarrengandik gertu daude, eta gauza bera gertatzen da “ardoa” eta “garagardoa” hitzekin. Bi bikote hauek elkarrengandik urrun daude. Adibide honen bitartez argi ikusten da gertutasun kontzeptu hori hitzen antzekotasun semantikoarekin lotuta dagoela. Word2Vec ereduarekin terminoen arteko gisa honetako erlazioak lor daitezke.

Eredu honen atzean bi arkitektura ezberdin daude [43]:

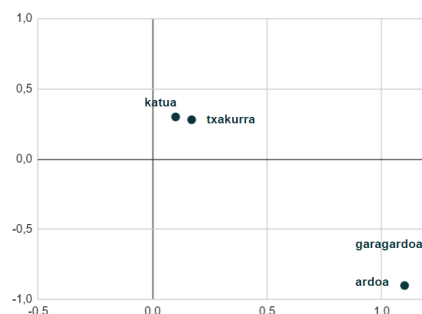
CBOW (Continuous Bag of Words). Eredu honen helburua testuinguru bat emanda bertan falta den hitza zein den iragartzea da. Esaterako, “Atzo gizona etxera joan zen.” Esalditik “zen” hitza kenduta eta ereduaren sarrera gisa gainontzeko hitzak erabilia, ereduaren helburua bertan falta den hitza zein den asmatzea litzateke.

²Kodea eskuragarri hemen: <https://github.com/gohjiayi/suicidal-text-detection/>

Continuous Skip-Gram Model. Eredu honek, aldiz, aurkakoa egiten du. Hitz bat eman da honek inguruan (aurretik eta ondoren) izan ditzakeen hitzak asmatzen saiatzen da.

	ardoa	garagardoa	katua	txakurra
x	1,1	1,2	0,1	0,17
y	-0,9	-0,7	0,3	0,28

4.3 Taula: 2-dimentsioko Word2Vec errepresentazioen adibideak, x eta y balioak hitz bakoitzaren XY espazioko koordenatuak izanik.



4.2 Irudia: 2-dimentsioko Word2Vec errepresentazioen adibideak grafikoki.

4.3.2.2 Topiko-ereduak

Bestalde, topiko-ereduak erabili dira hauek mezuen errerepresentaziorako izan dezaketen ahalmena aztertzeko asmoz. Topiko-eredua ikasketa automatikoko teknika da, dokumentu-bilduma batean egitura semantiko ezkutuak identifikatzeko gai dena [44, 45]. Estatistikan oinarritutako ereduak erabiliz, dokumentu-multzoak analizatzen ditu bertako hitz eta esaldien patroiak ikasteko asmoz [46]. Patroi horiek topiko deritzen hitz-multzoen bitartez errepresentatu ohi dira, corpusaren laburpen bat eskaintzen dutenak. Arrazoi hauengatik, teknika hau datu-bilduma ez-egituratueta nolabaiteko antolamendua aurkitzeko metodo baliagarri bilakatu da [47, 48].

Proiektu honetan, zehazki, Latent Dirichlet Allocation (LDA) erabili da topiko-ereduak lortzeko. Teknika hau erabilienetako bat da eta oinarritzat hiru mailako probabilitate-eredu bayestar hierarkiko bat du [44]. Topiko bakoitza hitzen gaineko probabilitate-banaketa gisa errepresentatzen du, eta dokumentu bakoitza topikoen banaketa gisa.

Konputazionalki, LDA topiko-eredua bi matrize gisa irudika daiteke, M_ϕ hitz-topiko matrizea eta M_θ dokumentu-topiko matrizea izanik. M_ϕ hitz-topiko matrizeak $W \times T$ dimentsioa du, W hiztegiko hitz kopurua eta T ereduaren topiko kopurua izanik. Matrize honen errenkada bakoitzean hiztegiko hitz batek topikoetako bakoitzean duen garrantzia gordetzen da (4.1. adierazpena). Hiztegiko n -garren hitzaren (w_n) topiko-errerepresentazioa honako bektorea litzateke: $\phi(w_n) = (\phi_{n,1}, \phi_{n,2}, \dots, \phi_{n,T})$, non $\phi_{n,i} = p(w_n|t_i)$ den. Matrize honen zutabe bakoitzean topiko bakoitzaren hitzen gaineko banaketa aurki daiteke. M_θ , aldiz, $D \times T$ dimentsioko matrizea da, D bildumako dokumentu kopurua delarik. Matrize honen errenkada bakoitzean bildumako dokumentu baten topiko-banaketa gordetzen da (4.2. adierazpena). Bildumako m -garren dokumentuaren (d_m) topiko-errerepresentazioa honako bektorea litzateke: $\theta(d_m) = (\theta_{m,1}, \theta_{m,2}, \dots, \theta_{m,T})$, non $\theta_{m,i} = p(t_i|d_m)$ den [49].

$$M_\phi = \begin{bmatrix} \phi_{11} & \phi_{12} & \dots & \phi_{1T} \\ \phi_{21} & \phi_{22} & \dots & \phi_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{W1} & \phi_{W2} & \dots & \phi_{WT} \end{bmatrix} \quad (4.1)$$

$$M_\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1T} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{D1} & \theta_{D2} & \dots & \theta_{DT} \end{bmatrix} \quad (4.2)$$

Hainbat ikerketetan ikusi da dokumentu-bilduma baten gainean topiko-ereduak aplikatzean dokumentu bakoitzarentzat lortutako probabilitate-banaketak honen dimentsionalitate baxuko errepresentazioa eskaintzen duela [50, 51]. Proiektu honetan hau erabili da sare sozialetako mezuen errepresentazio alternatibo gisa.

LDA ereduaren implementatu eta errepresentazioak lortzeko Python lengoaiako gensim liburutegia erabili da.

4.3.3 Sailkatzaileak

Mezuen sailkapena egin ahal izateko beharrezkoa da sailkatzaileak erabiltzea. Hauen funtzioa sarrera gisa erabiliko mezuen errepresentaziotik abiatuta etiketen erlazioak ikasi eta mezuei dagokien etiketa aurreratu da.

Proiektu honetan hainbat sailkatzaile ezberdin erabili dira, bi multzo ezberdinetan bereizten direnak: ikasketa automatikoan oinarritutakoak eta ikasketa sakonean oinarritutakoak. Horietaz gain, LDA topiko-eredutik abiatuta lortutako errepresentazioentzat oinarri-lerro gisa erabiliko den sailkatzaile bat proposatu da lan honetarako espresuki. Jarraian eredu hauek guztiak azaltzen dira.

4.3.3.1 LDA topiko-ereduarentzat proposatutako oinarri-lerroa

Proiektu honetan, LDA topiko-ereduaz baliatuz lortutako errepresentazioentzat oinarri-lerro gisa erabiliko den sailkatzaile bat proposatzea erabaki da. Eredu honetan, \mathbf{x}_i mezu bakoitzaren topiko-errepresentazioan gailentzen den topikoari ematen zaio garrantzia. Topiko hau $tGailendu(\mathbf{x}_i)$ bezala adierazi da (4.3) adierazpenean. Izan bedi $\mathcal{X}(t_j, c_k)$ entrenamenduko instantzien azpimultzoa halakoa non denetan gailentzen den topikoa t_j den eta gainera mezu guztiak c_k klasekoak diren. Izan bedi $tcModa(t_j)$ topiko nabarmenena t_j duten instantzien arteko modako klasea (maizen agertzen den etiketa), (4.5) adierazpenean adierazi den moduan.

Oinarri-lerroaren entrenamendu fasean topiko bakoitzeko modako klasea gordetzen da: $t_j \rightarrow tcModa(t_j)$. Hau da, topiko bakoitzeko, topiko hori gailentzen den entrenamenduko instantzien azpimultzoko modako klasea gordetzen da. Ereduaren ustiapen fasean, mezu berri bat emanda, testua topikoen espazioan adierazten da lehendabizi (\mathbf{x}_{berria}). Ondoren, oinarri-lerroak bere klasea iragartzen du (4.6) adierazpenean azaltzen den moduan. Hitz gutxitan esanda, instantziaren errepresentazioan gailentzen den topikoarentzat entrenamendu fasean gordetako modako klasea esleitzen zaio adibide berriari.

$$tGailendu(\mathbf{x}_i) = t_s \quad : \quad s = \arg \max_{1 \leq j \leq m} t_{ij} \quad \wedge \quad \mathbf{x}_i = (t_{i1}, t_{i2}, \dots, t_{im}) \quad (4.3)$$

$$\mathcal{X}(t_j, c_k) = \{(\mathbf{x}_i; c_i) \in X_{Train} | tGailendu(\mathbf{x}_i) = t_j \wedge c_i = c_k\} \quad (4.4)$$

$$tcModa(t_j) = \arg \max_{c_k \in \mathcal{C}} |\mathcal{X}(t_j, c_k)| \quad (4.5)$$

$$\hat{c}(\mathbf{x}_{berria}) = tcModa(tGailendu(\mathbf{x}_{berria})) \quad (4.6)$$

4.3.3.2 Ikasketa automatikoan oinarritutakoak

Ikasketa automatikoko sailkatzaile klasikoek algoritmoak erabiltzen dituzte datuetatik era automatikoan ikasteko eta erabakiak hartzeko. Sailkatzaile klasiko hauek, ordea, datu gordinak tratatzeko gaitasun mugatua dute, eta ondorioz, ezaugarri iragarleen aukeraketaren gaineko lanketa eskatzen dute. Kasu honetan, ikasketa automatikoan oinarritutako hiru meta-sailkatzaile erabili dira.

Random Forest. Sailkapen-zuhaitzetan oinarritzen den meta-sailkatzailea da. Bagging teknika du oinarri, zuhaitz anitzen iragarpenak konbinatzen ditu elementu baten klasea iragartzeko [52].

Bagging. Meta-sailkatzaile honen bitartez oinarritzko sailkatzaile bera indusitzen da entrenamendurako azpimultzo ezberdinetatik. Behin guztiak indusitu direla, bakoitzaren iragarpenen konbinaketa egiten da azken iragarpena lortzeko [53]. Kasu honetan, oinarritzko sailkatzaile gisa C-Support Vector Classifier (C-SVC) erabili da.

XGBoost. Boosting teknika erabiltzen duen metodoa da hau, abiapuntu gisa erabakizuhaitzak dituena. Random Forest bezalako teknikek baino emaitza hobekak eman ohi ditu [54].

Sailkatzaile klasikoen ereduak lortzeko Python lengoaiako sklearn liburutegia erabili da.

4.3.3.3 Ikasketa sakonean oinarritutakoak

Bestalde, ikasketa sakonean oinarritutako sailkatzaileak daude. Hauek ikasketa automatikoko sailkatzaile aurreratu gisa uler daitezke, gai baitira datuetatik ezaugarri iragarleak automatikoki ateratzeko eta erlazio konplexuagoak ikasteko. Horregatik, ikasketa automatikoko ereduak baino emaitza hobekak lortu ohi dituzte eta egungo artearen egoera osatzen dute [55]. Hala ere, eredu hauen tamaina dela-eta, makina-baliabide kantitate handiak behar dituzte eta hauek entrenatzeak dakarren zama konputazionala sailkatzaile klasikoena baina nabarmenki handiagoa da. Lan honetan artearen egoerako bost eredu ezberdin landu dira, *Suicidal-BERT* lanean erabilitakoak erreproduzitzen. Jarraian, bakoitzaren ezaugarri orokorrak azaltzen dira.

Erregresio logistikoa. Eredu hau ikasketa sakonean oinarritzen ez den arren, hizkuntzaren prozesamenduan sailkapenerako oinarri-lerro gisa erabili ohi den metodo estatistikoa da [56].

Neurona-sare konboluzionala (CNN). Neurona-sare mota hau konboluzioetan oinarritzen da, datuetatik ezaugarri edo patrioiak erauzteko gai direnak [57]. Ikusmen artifizialean erabiltzen dira gehienbat, baina badaude testuekin lan egiteko prestatutako ereduak ere.

LSTM neurona-sarea. Neurona-sare errekurrente (Recurrent Neural Network, RNN) mota bat da, desagertzen diren gradienteen arazoa konpontzen duena [58]. Testu, audio edota bideo moduko datu sekuentzialetan erabiltzen dira gehien bat.

BERT transformerra. Googlek sortutako eredu da, atentzio-mekanismoa erabiltzen duena [59]. Gaur egun oso erabilia da hizkuntzaren prozesamenduan eta tamaina ezberdinetako bertsiok daude eskuragarri.

ELECTRA transformerra. Hau ere Googlek sortutako transformer eredu da [60]. BERTekin alderatuta, entrenamendu datu gutxiago behar ditu errendimendu bera lortzeko, adibide bakoitzeko informazio gehiago jasotzeko gaitasuna baitu.

4.3.4 Ebaluazioa

Eredu ezberdinek egindako iragarpenen kalitatea neurtzea ezinbestekoa da hauek konparatu ahal izateko. Jarraian, erabili diren ebaluazio-metrikak eta ebaluazio-eskema azaltzen dira.

4.3.4.1 Ebaluazio-metrikak

Sailkatzaile ezberdinekin lortutako emaitzak ebaluatzeko lau metrika erabili dira; metrika hauen adierazpen matematikoak (4.7)-(4.10) adierazpenetan aurki daitezke. Hauetan, lau gai erabiltzen dira: EP (egiazko-positibo), EN (egiazko-negatibo), PF (positibo-falstu) eta NF (negatibo-falstu). Gure lanean, positibo gisa hartutako etiketa *suicide* da.

$$\text{Asmatze-tasa} = \frac{EP + EN}{EP + EN + PF + NF} \quad (4.7)$$

$$\text{Doitasuna} = \frac{EP}{EP + PF} \quad (4.8)$$

$$\text{Estaldura} = \frac{EP}{EP + NF} \quad (4.9)$$

$$\text{F puntuazioa} = \frac{2 \cdot \text{doitasuna} \cdot \text{estaldura}}{\text{doitasuna} + \text{estaldura}} \quad (4.10)$$

Metrika hauez gain, kasu batzuetan nahasmen-matrizea eraiki da, non EP, EN, PF eta NF kopuruak adierazten diren taula batean laburbilduta.

4.3.4.2 Ebaluazio-eskema

Ebaluazio-eskema bezala *hold-out* erabili da. Datuen prestaketa egin ostean lortutako datu-sorta hiru azpimultzotan banatu da klase-banaketa errespetatuz: entrenamenduko azpimultzoa, ereduak entrenatzeko erabiltzen dena (%80); garapeneko, entrenamendu prozesuan ereduak ebaluatzeko erabiltzen dena (%10); eta probakoa, entrenamenduaren ostean ereduak ebaluatzeko erabiltzen dena (%10). Eredu guztietan partiketa bera erabili da, bakoitzarekin lortutako emaitzak ahalik eta konparagarrienak izateko.

Egindako partiketa ez da aurrekariak erabilitako bera izan. Honen arrazoia datu-sorta hau erabiltzen duten lanek partiketa ezberdinak erabili dituztela izan da, eta kasu batzuetan datu-sorta osoa erabili ordez azpimultzo bat soilik erabili dela (ikus [4.1.2.](#) atala).

Mezuen analisi linguistikoa

Kapitulu honetan mezuen analisi linguistikoaren inguruan egindako lana azaltzen da. Aurrekarien inguruan hitz egiten den 3.2.1. atalean azaldu bezala, hainbat lanek jorratu dute gai hau eta bakoitzak ezaugarri ezberdinak aztertu ditu, ondorio ezberdinak ateraz. Proiektu honetan horietako batzuk aukeratu eta denak datu-sorta beraren gainean konprobatzea erabaki da, bertan ere betetzen diren ikusteko asmoz. Hautatutako ezaugarriak mezuen luzera, galdera-marka kopurua, lehen pertsona singularreko izenordainen presentzia eta aditz-denboren erabilera izan dira.

Lan honetan erabilitako *Suicide and Depression Detection* datu-sortaren gainean orain arte ez da mezuen analisi linguistiko esanguratsurik egin, 4.1. atalean azaldu bezala. Beraz, proiektu honetan egindako azterketaren bitartez bide honi hasiera eman nahi izan zaio.

Kapitulua bost ataletan banatzen da. Lehen lau ataletako bakoitza landutako ezaugarri bati dagokio. Ezaugarri bakoitzeko, lehenbizi hau analizatzeko jarraitutako prozedura azaltzen da eta ondoren analisi horretatik eratorritako emaitzak aztertzen dira. Azken atalean, egindako analisisian lortutako emaitzak laburbiltzen dira.

5.1 Mezuen luzera

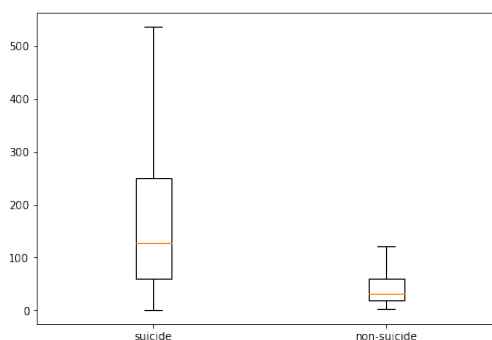
Lehenik eta behin datu-sortako mezuen luzera aztertu nahi izan da. Izan ere, aurrekarien analisisian aipatu den bezala, suizidio-ideiagintzarekin erlazionatutako txioak luzeagoak direla ikusi da [24]. Kasu honetan, ezaugarri hau aztertzeko `split()` eta `len()` funtzioak erabili dira, 4.2.1.1. atalean azaldu direnak.

Klaseen arteko bereizketa oso argia dela ikus daiteke 5.1. taulan eta 5.1. irudiko kutxa-bibote diagraman laburbildu diren emaitzetan. Suizidio-ideiagintzaren zantzuak erakusten dituzten mezuen batez besteko luzera 203 hitzekoa da, gainontzeko mezuek 61 hitzekoa den bitartean. Kuartilei dagokienez, *non-suicide* etiketadun mezuen % 75a 60 hitzetik beherakoa da, *suicide* kasuan 60 hitzetik beherako mezuen portzentaia % 25ekoa den bitartean. Hau oso garbi ikusten da diagraman, bi kutxak ez baitira puntu batean ere gainjartzen.

Mezuen luzera atributu iragarle gisa erabiliko balitz, 60 hitzetik gorako mezuei *suicide* etiketa esleituz eta aurrakako kasuan *non-suicide* klaseko mezutzat hartuz, % 74,7ko asmatzetasia lortuko litzateke. Honek, printzipioz, hasierako hipotesia baieztatzen du, hau da,

	<i>suicide</i>	<i>non-suicide</i>
\bar{x}	203	61
σ	255	139
min	1	2
% 25	60	19
% 50	127	31
% 75	251	60
max	9.684	8.220

5.1 Taula: Mezuen luzerarekin erlazioatutako estatistikoak klaseka.



5.1 Irudia: Mezuen luzeraren kutxa-bibote diagrama klaseka.

Twitterren bezala, Reddit sare sozialeko mezuak oinarri dituen datu-sorta honetan ere ideia gintza suizidarekin lotutako mezuak luzeagoak direla.

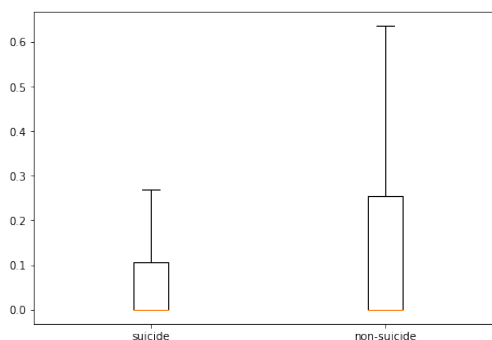
5.2 Galdera-markak

Beste alde batetik, suizidio-ideia gintzarekin lotutako Redditeko mezuetan galdera-marken presentzia handiagoa denaren ebidentzia aurkitu da [26]. Hau proiektu honen ardatz den datu-sortaren gainean frogatzeko, mezu bakoitzean ‘¿’ eta ‘?’ ikurren gaineko bateragarritasun zehatz kopurua zenbatu da.

Lortutako emaitzak 5.2. taulan eta 5.2. irudiko diagraman laburbildu dira. Taulan ikus daitekeenez *non-suicide* 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-ideia gintzarekin erlaziorik ez duten mezuen kasuan aldakortasuna handiagoa dela. Maiztasun absolutuak aztertuz, bi kasuetan mezuek galdera-marka bat baino gutxiago erabiltzen dute. Aipatu beharrekoa da *non-suicide* 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.

	<i>suicide</i>		<i>non-suicide</i>	
	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

5.2 Taula: Mezuen galdera-marka kopuruarekin erlazioatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.



5.2 Irudia: Mezuen galdera-marka kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatuak datuetatik abiatuta eraikia.

Beraz, proiektu honetan 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 *non-suicide* gisa etiketatutako mezuek gehiago erabiltzen dituztela ikusi da.

5.3 Lehen pertsona singularreko izenordainak

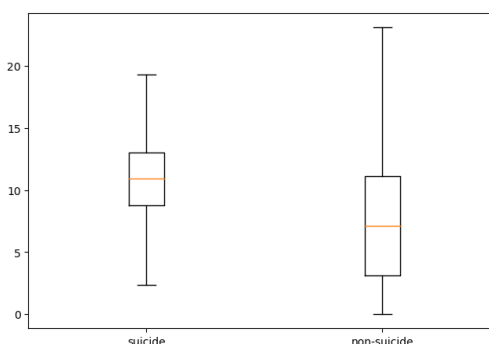
Hainbat lanek ondorioztatu dute suizidio-ideiagintzaren zantzuak erakusten dituzten testu mota ezberdinetan ohi baino lehen pertsona singularreko izenordain gehiago erabiltzen direla [22, 23, 24, 25, 26]. Ezaugarri linguistiko hau aztertzeko, UDPipe tresnaren (informazio gehiago 4.2.1.3. atalean) XPOS (hizkuntzaren menpeko *part-of-speech*) gramatika-kategoriaren etiketatzailea erabili da. 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.

Datu-sortako mezu guztiak aztertu ostean, lortutako emaitzak 5.3. taulan eta 5.3. irudiko diagraman laburbildu dira. Kasu honetan, bi etiketen arteko ezberdintasuna nabaria da maiztasun absolutuko datuak behatuz. Hirugarren kuartilari erreparatuz, esaterako, ikus daiteke *non-suicide* etiketadun mezuen laurden batek soilik erabiltzen dituela 5 izenordain baino gehiago, *suicide* etiketadunen hiru laurdenek 7 baino gehiago erabiltzen dituzten bitartean.

Normalizatutako datuen kasuan, nahiz eta ezberdintasun hori presente dagoen, ez da hain nabaria. *suicide* etiketadun mezuen kasuan lehen pertsona singularreko izenordainek batez beste hitzen % 10,7a hartzen dute, *non-suicide* etiketadunen kasuan hitz hauen batez besteko portzentaia % 7,39 den bitartean. Diagrama behatuz ikusten da suizidio-ideiagintzarekin erlazionatutako mezuetan izenordain hauen erabilera nabarmenagoa dela, medianaren diferentzia ia lau puntukoa izanik. Hala ere, *non-suicide* 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.

	<i>suicide</i>		<i>non-suicide</i>	
	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	696	64,2

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



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

Beraz, 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.

5.4 Aditzen denborak

Hainbat lanetan, suizidio-ideiagintzarekin erlazionatutako testuetan denborarekin lotutako ezaugarriak identifikatu dira. Horietako batean [23], esaterako, suizidio-oharretan geroaldiko aditz gutxiago erabiltzen direla ondorioztatu zen, eta txioak aztertzen zituen beste batean [24] ideia gintza suizida presente dagoen txioetan orainaldian fokua jartzen dela.

Modu ezberdinak daude denborazkotasuna 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.

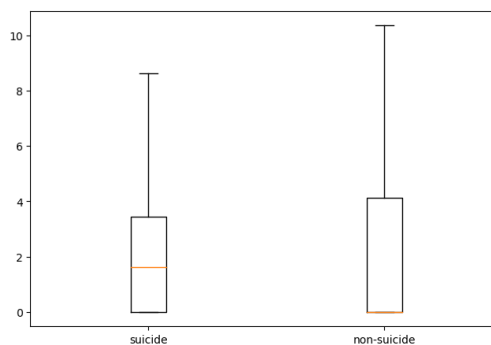
5.4.1 Lehenaldia

Izenordainekin egin den antzera, lehenaldiko aditzak identifikatzeko gramatika-kategoriaren etiketatzailea erabili da. Izan ere, badago aditz hauei esleitzen zaien XPOS etiketa bat, VBD (*Verb, past tense*). Beraz, UDPipe erabiliz hitz bakoitzaren XPOS etiketa aztertu da eta mezu bakoitzeko lehenaldiko aditz kopurua gorde de.

Emaitzak laburbiltzen dituzten 5.4. taulan eta 5.4. irudiko diagraman ikus daiteke bi klaseetan lehenaldiko aditzen antzeko erabilera egiten dela. Ezberdintasun nagusia medianan dago, *suicide* etiketadun mezuen kasuan handiagoa baita.

	<i>suicide</i>		<i>non-suicide</i>	
	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

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



5.4 Irudia: Mezuen lehenaldiko aditz kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.

Hortaz, badirudi datu-sorta honetan lehenaldiko aditzak era berean erabiltzen direla suizidio-ideiagintzarekin erlazionatutako mezuetan eta honekin zerikusirik ez dutenetan.

5.4.2 Orainaldia

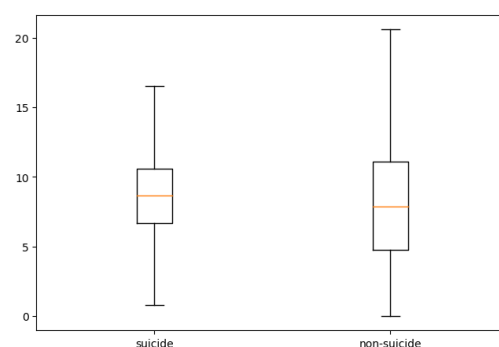
Orainaldiko aditzekin lehenaldikoekin egin den antzera jokatu da, XPOS etiketak erabiliz. Kasu honetan, bi etiketa daude orainaldiko aditzentzat: VBP (*Verb, non-3rd person singular present*), hirugarren pertsona singularrari erreferentzia egiten ez dioten orainaldiko aditzentzat eta VBZ (*Verb, 3rd person singular present*), hirugarren pertsona singularreko

orainaldiko aditzei dagokiena. Beraz, hitz bakoitzaren etiketa aztertu da aipatutako bien bila, eta mezua bakoitzerako hauen agerpen kopurua gorde da.

Emaitzak behatuz (5.5. taula eta 5.5. irudiko diagrama), luzerarekiko normalizatutako datuetan ageri den ezberdintasun nagusia datuen dispersioan dago. Hau handiagoa da *non-suicide* etiketadun mezuen kasuan, diagramako kutxetan ikusten denez. Suizidio-ideiagintzarekin erlazioatutako 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.

	<i>suicide</i>		<i>non-suicide</i>	
	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

5.5 Taula: Mezuen orainaldiko aditz kopuruarekin erlazioatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.



5.5 Irudia: Mezuen orainaldiko aditz kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.

Beraz, esan daiteke bi kasuetan orainaldiko aditzen antzeko erabilera egiten dela, *non-suicide* etiketadunen kasuan aldakortasuna handiagoa dela kontuan izanik.

5.4.3 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 [61]. Etorkizunari erreferentzia egiteko erabili ohi den beste egitura bat “*be going to*” da.

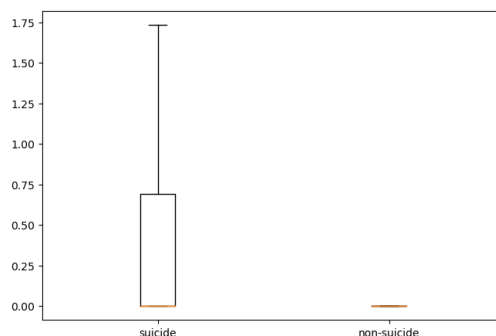
Ikusten denez, gai eztabaidatsua da geroaldiarena. Lan honetarako, aipatutako bi egiturak izan dira kontuan, “*will*” eta “*be going to*”. Lehenengoaren kasuan, lema gisa “*will*” hitza duten tokenak bilatu dira, MD (*modal*) XPOS etiketa dutenak, eta dependetziei dagokienez, *head* edo burutzat aditz 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.

Bi egitura hauen bilaketaren emaitzak 5.6. taulan eta 5.6. irudiko kutxa-bibote diagraman laburbildu dira. Luzerarekiko [0-100] eskalan normalizatutako batezbesteko eta desbideratze estandarri erreparatuta esan daiteke bi klaseek aditz-denbora honen antzeko

erabilera egiten dutela, eta baita presentzia gutxi dutela ere. Maiztasun absolutuko datuei dagokienean, *suicide* klaseko mezuetan batez beste etorkizunari erreferentzia egiten dion aditz bat erabiltzen dela ikusten da, *non-suicide* klasekoetan kopuru hori zerora hurbiltzen den bitartean. Diagraman, *non-suicide* 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.

	<i>suicide</i>		<i>non-suicide</i>	
	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

5.6 Taula: Mezuen geroaldiko aditz kopuruarekin erlazioatutako estatistikoak klaseka, maiztasun absolutuak (abs.) eta luzerarekiko [0-100] eskalan normalizatuak (norm.) bereiziz.



5.6 Irudia: Mezuen geroaldiko aditz kopuruaren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.

Hortaz, esan daiteke bi klaseetan etorkizunari erreferentzia gutxi egiten zaiola, aztertutako egiturei dagokienean behintzat. Hala ere, hau apur bat presentego dago suizidio-ideiagintzarekin erlazioatutako mezuetan.

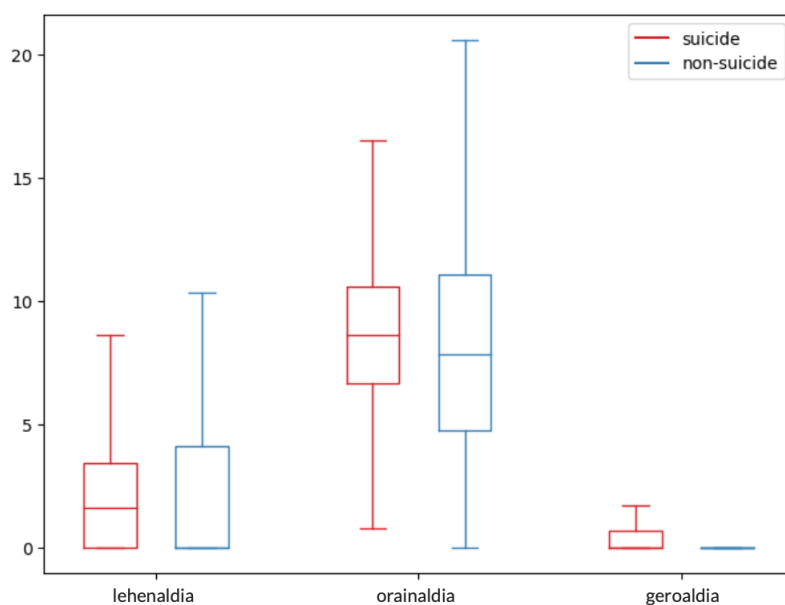
5.4.4 Hiru aditz-denboren arteko konparaketa

Hiru aditz-denborak era isolatuan aztertu ostean, beraien arteko konparaketa egitea erabaki da. Horretarako, 5.7. irudiko kutxa-bibote diagrama eraiki da, non aditz-denbora bakoitzari dagokion grafikoa irudikatu den bi klaseak koloretan bereiziz eta luzerarekiko [0-100] eskalan normalizatutako datuak erabiliz.

Bi klaseetan diferentzia esanguratsurekin 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 *non-suicide* etiketadun mezuen kasuan eta ondorioz, hau ezin da suizidio-ideiagintzarekin erlazioatutako mezuen ezaugarri bereizgarritzat hartu.

Bestalde, etorkizunari erreferentzia egiten dioten geroaldiko aditzei dagokienean, ikusten da presentzia gutxi dena dutenak direla bi kasuetan. Suizidio-ideiagintzaren zantzuak erakusten dituzten mezuen kasuan presentzia hori apur bat nabarmenagoa da, nahiz eta diferentzia oso txikia izan. Beraz, emaitza hauek ikusita, ezin daiteke esan ideia gintza suizidaren zantzuak dituzten mezuek etorkizunari erreferentzia gutxiago egiten diotenik.

Laburbilduz, *Suicide and Depression Detection* datu-sortan ez dira betetzen lan honetan aztertu diren eta aditz-denborekin erlazioatuta dauden bi ezaugarriak.



5.7 Irudia: Mezuen aditz-denboren erabileraren kutxa-bibote diagrama klaseka, luzerarekiko [0-100] eskalan normalizatutako datuetatik abiatuta eraikia.

5.5 Analisiaren laburpena

Aztertutako lau ezaugarriekin lortutako emaitzak ikusita ondorioztatzen da *Suicide and Depression Detection* datu-sortan suizidio-ideiagintzaren zantzuak erakusten dituzten mezuak luzeagoak direla eta lehen pertsona singularreko izenordain gehiago erabiltzen dituztela, aurrekariak adierazitakoarekin bat etorriz [22, 23, 24, 25, 26].

Bestalde, ikusi da kasu honetan ideia gintza suizidarekin erlazionatutako mezuetan galdera-markek duten presentzia ez dela nabarmentzekoa, bi klaseek antzeko maiztasunez erabiltzen baitituzte. Aditz-denboren inguruan egindako azterketak emandako emaitzek ere erakutsi dute ez dagoela klaseen arteko ezberdintasun nabarmenik.

Mezuen sailkapena

Mezuen gaineko sailkapen bitarrari egiten zaio aurre kapitulu honetan. Ataza honi dagokion artearen egoeraren azterketan ikusi denez (3.2.2. atala), metodo ezberdinak erabili izan dira hau egiteko. Proiektu honetan, metodo ezberdin horietatik abiatuta bi hurbilpen ezberdin diseinatu eta inplementatu dira, bata artearen egoerako sistemek osatua eta bestea oinarrizkoagotzat jotzen diren ereduak lortua. Honen helburua bi hurbilpenak konparatzea izan da, eta oinarrizkoagoa den hurbilpen horrekin artearen egoerako emaitzetara gerturatzea posible den ikustea.

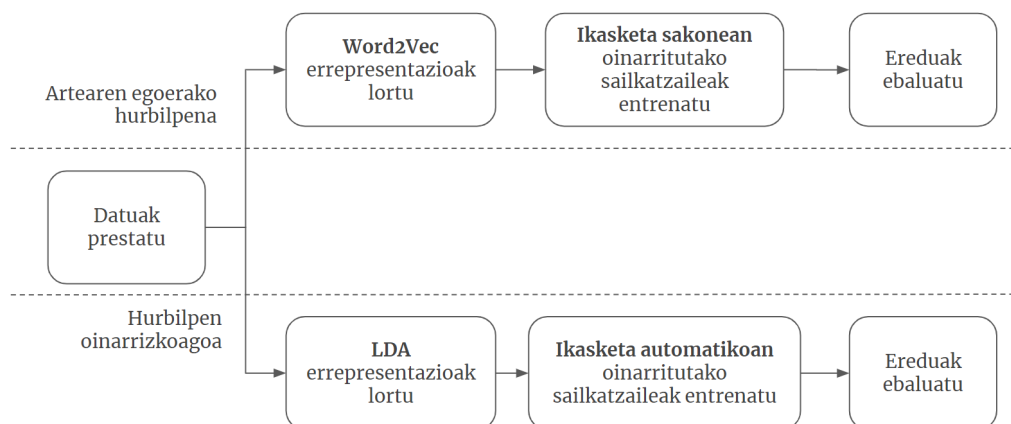
Kapitulu honetan, lehendabizi esperimentuen diseinua azaltzen da, non hurbilpen bakoitza azaltzen den modu zehatzagoan. Ondoren, datuen gainean egindako prestaketa prozesuari buruzko xehetasunak adierazten dira, eta jarraian, aukeratutako bi hurbilpenen inplementazioaren inguruko zehaztasunak eta bakoitzarekin lortutako emaitzak aurkeztu eta eztabaidatzen dira. Amaitzeko, esperimentaziotik ateratako ondorio nagusienak laburbiltzen dira azken atal batean.

6.1 Esperimentuen diseinua

Esan bezala, mezuen gaineko sailkapen bitarra egiteko bi hurbilpen landu dira proiektu honetan. Alde batetik, **artearen egoerako hurbilpena** deritzona eraiki da. Hurbilpen honetan, izenak dioen bezala, ikerketa arlo honetan artearen egoera osatzen duten sistemak erabiltzen dira: mezuak numerikoki errepresentatzeko Word2Vec hitz-embeddingak entrenatu dira lan honetan erabiltzen den datu-sortaren gainean, eta ondoren ikasketa sakonean oinarritutako sailkatzaile ezberdinekin sailkatu. Bestetik, **hurbilpen oinarrizkoagoa** sortu da, artearen egoerakoarekin lortutako emaitzetara hurbiltzeko asmoz. Honetan, mezuen errepresentazio modu alternatibo bat proposatu da, topiko-ereduak erabiliz lortua, eta hau sarrera gisa hartuta ikasketa automatikoan oinarritutako sailkatzaileak erabili dira mezuak sailkatzeko.

Bi hurbilpen horietarako, materialak eta metodoak kapituluko 4.3. atalean ezarritako 4 pausoak jarraitu dira: datuak prestatu, mezuen errepresentazioa lortu, sailkatzaileak entrenatu eta ereduak ebaluatu. Pauso bakoitza gauzatzeko egindako lana, ordea, ezberdina izan da, hurbilpen bakoitzean sistema ezberdinak erabiltzen direlako. Esperimentazioan

jarraitutako eskema 6.1. irudiko diagraman aurki daiteke, non hurbilpen bakoitzerako jarraitutako prozedura azaltzen den.



6.1 Irudia: Mezuen sailkapena egiteko jarraitutako esperimentazio eskema.

Lehendabizi datuen prestaketa egin da sare sozialetako testu desegituratua nolabait txukuntzeko eta ondoren eraikiko diren ereduaren entrenamendu prozesua arintzeko asmoz. Prestaketa hau komuna izan da bi hurbilpenetarako. Ondoren, hurbilpen bakoitzeko errepresentazio numerikoak lortu dira, artearen egoerako hurbilpenearan Word2Vec hitz-embeddingak erabiliz eta hurbilpen oinarritzkoagoaren kasuan LDA topiko-ereduarekin. Ondoren, hurbilpen bakoitzerako aukeratutako sailkatzaileak entrenatu dira: lehenengoaren kasuan ikasketa sakonean oinarritutakoak eta bigarrean aldiz, ikasketa automatikoa oinarri dutenak. Amaitzeko, ereduak ebaluatu dira 4.3.4. atalean azaldutako prozedura jarraituz.

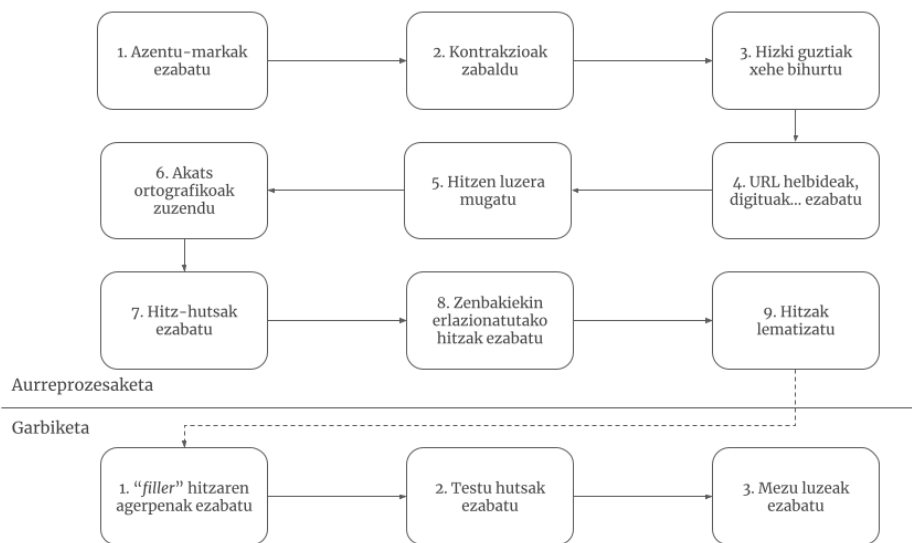
6.2 Datuen prestaketa

Egindako lehen pausoa datuak prestatzea izan da. Prozesu hau bi fasetan banatu da, aurreprozesaketa eta garbiketa. Fasetako bakoitzean hainbat urrats eman dira helburu ezberdinekin. Hauek guztiak 6.2. irudiko diagraman¹ laburbildu dira. Jarraian horietako bakoitza azaltzen da modu zehatzagoan. Ondoren, prestaketa prozesuak datu-sortan izandako eragina eta sailkapenerako gauzatutako partiketa deskribatzen dira.

6.2.1 Aurreprozesaketa

Aurreprozesaketan hainbat urrats ezberdin eman dira. Hauen helburu nagusia hiztegiaren tamaina txikitzea izan da, datuen dimentsionaltasuna murrizteko eta ereduaren entrenamendu prozesuaren zama konputazionala arintzeko. Gainera, emandako urratsek mezuetan egon daitekeen zarata minimizatzen eta ataza honetarako garrantzitsua ez den informazioa ezabatzen lagundu dute. Emandako pausuak honakoak izan dira:

¹Diagrama hau sortzean *Suicidal-BERT* proiektuko autoreek egindakoa hartu da erreferentzia gisa. https://github.com/gohjiayi/suicidal-text-detection/blob/main/docs/Suicidal-Text-Detection_Report.pdf



6.2 Irudia: Datuen prestaketaren laburpena, aurreprozesaketa eta garbiketaren jarraitutako prozesua adieraziz, pausoz pauso.

1. Azentu-markak ezabatu: testuinguru honetan, “*café*” eta “*cafe*” hitzek esanahi bera dutenez azentu-markak ezabatu egin dira.
2. Kontrakzioak zabaldu: ingelesez baliabide morfologiko ohikoa da kontrakzioa, bi hitzen elkarketan datzana. Datuak estandarizatzeko, kontrakzioak zabaltzea erabaki da.
3. Hizki guztiak xehe bihurtu: ataza honetan letra larriz idatzitako hitzek ez dute arreta berezirik behar. Datuen dimentsionaltasuna murrizteko hizki guztiak xehe bihurtu dira.
4. URL helbideak, sinboloak, digituak, karaktere bereziak eta zuriune estrak ezabatu: lan honen helburuan hauek guztiek garrantzia berezirik ez dutenez, ezabatzea erabaki da.
5. Hitzen luzera mugatu: datu-sorta honetako hainbat mezuetan era okerrean errepikatutako hizkiak aurkitu dira, “*goood*” bezalakoak, berez “*good*” izango litzatekeena. Errepikapenak bira murriztea erabaki da, ingelesez hau baita jarraian aurki daitekeen hizki beraren agerpen kopurua.
6. Akats ortografikoak zuzendu: sare sozialen tankeko testuinguru informalean ez zaio hainbesteko garrantzia ematen zuzentasun ortografikoari. Aurki daitezkeen akatsak zuzentzen saiatzeko asmoz, Pythoneko *Symspell* algoritmoa erabili da.
7. Hitz-hutsak (*stopwords*) ezabatu: funtzio-hitz asko hitz-huts kontsideratzen dira, lokailuak esaterako. Ataza honetan fokua eduki-hitzetan jarriko denez, hauek ezabatzea erabaki da. Ezeztapena adierazten duten “*no*” eta “*not*” hitzak askotan gisa honetakoak kontsideratzen diren arren, ataza honetan negatibotasuna esanguratsua izan daitekeenez, mantendu egin dira.

8. Zenbakiekin erlazionatutako hitzak ezabatu: digituak ezabatu diren arren, batzuetan zenbakiak hitzez adierazten dira, “bost” esaterako.
9. Hitzak lematizatu: lematizazioa analisi morfologikoaren bidez hitzen lemak eta morfemak lortzeko prozesua da. Hitz guztiak beren lemengatik ordezkatu dira.

6.2.2 Garbiketa

Aurreprozesaketarekin lortutako emaitzak txukuntzeko asmoz, garbiketa egin da. Kasu honetan, pauso hauek jarraitu dira:

1. “*filler*” hitzaren agerpenak ezabatu: hitz honek agerpen kopuru handia izan arren, ondorioztatu da bere jatorria datuak eskuratzeko prozesuan sortutako zarata izan daitekeela. Honengatik, ezabatzea erabaki da.
2. Testu hutsak ezabatu: aurreprozesaketaren ostean testu asko hutsak bilakatu direnez, ezabatu egin dira.
3. Mezu luzeak ezabatu: jarraian azalduko den 6.1a. taulari erreparatuta, ikus daiteke aurreprozesaketaren ostean mezu luzeena 5.850 hitzekoa dela, 75. pertzentila 61 hitzekoa izanik. Entrenamendu prozesua arintzeko asmoz, 62 hitzetik gorako mezuak ezabatzea erabaki da.

6.2.3 Prestaketaren eragina

Prestaketa prozesuak datu-sortan izan duen eragina laburbiltzen du 6.1. taulak. Zehazki, bi azpitaulari sortu dira: lehenengoan (6.1a. taula), mezuen luzerarekin (hitz kopuruarekin) lotutako estatistiko ezberdinak aurki daitezke: batezbestekoa (\bar{x}), desbideratze estandarra (σ), kuartilak (% 25, % 50 eta % 75) eta luzera minimo eta maximoak (min eta max). Bigarrenean (6.1b. taula), mezuen klase-banaketa dago bilduta. Bi azpitauletan datuak hiru zutabetan banatzen dira, bakoitza prestaketako fase bati dagokiolarik: jatorrizko datu-sorta; aurreprozesaketa osteko datu-sorta; garbiketa osteko datu-sorta.

Aurreprozesaketa egin ondoren mezuen luzeraren batezbestekoak eta desbideratze estandarrak beherakada nabarmena izan dute. Garbiketak ere izan du eragina mezuen hitz kopuruan. Adibidez, hasiera batean mezu batek batez beste 132 hitz zituen eta aurreprozesaketaren eta garbiketaren ondoren, balio hau 21,4 hitzetan geratu da. Hiztegia txikitu da eta informazioa trinkotu egin da. Gainera, garbiketak aldaketak eragin ditu klase-banaketan. Izan ere, garbiketarekin, aurretik zegoen banaketa uniforme hautsi da, *non-suicide* klaseko mezu portzentaia datu-sorta osoaren % 61,1koa izanik, *suicide* klasekoena % 38,9koa den bitartean.

6.2.4 Partiketa

Sailkapena egiteko ebaluazio-eskema gisa *hold-out* metodoa erabili da, 4.3.4.2. atalean azaldu den bezala. Datu-sorta hiru azpimultzotan banatu da, prestaketa ostean datu-sortak izan duen klase-banaketa errespetatuz: entrenamendu-azpimultzoa (mezuen % 80), garapen-azpimultzoa (mezuen % 10) eta proba-azpimultzoa (mezuen % 10). Datu-sortaren eta zehazki egindako partiketaren aberastasun lexikoaren berri izateko asmoz honen analisi txiki bat egin da eta lortutako emaitzak 6.2. taulan laburbildu dira.

	Datu-sorta	Aurreproz.	Garbiketa
\bar{x}	132	52	21,4
σ	217	87,5	15,5
min	1	1	1
% 25	26	11	9
% 50	60	25	16
% 75	155	61	31
max	9.684	5.850	62

(a) Mezuen hitz kopuruaren bilakaera.

	Datu-sorta	Aurreproz.	Garbiketa
<i>suicide</i>	116.037 (% 50)	116.037 (% 50)	68.396 (% 38,9)
<i>non-suicide</i>	116.037 (% 50)	116.037 (% 50)	107.429 (% 61,1)
mezu kop.	232.074 (% 100)	232.074 (% 100)	175.825 (% 100)

(b) Mezuen klase-banaketaren eta datu-sortako mezu kopuruaren bilakaera.

6.1 Taula: Datu-sortaren deskribapena prestaketaren fase bakoitzean.

Emaitzak behatuz, azpimarratzekoa da entrenamendu-multzoko hiztegiaren tamaina, 27.523 hitzekoa dena. Are eta gehiago, hapaxek, azpimultzoan agerpen bakarra duten hitzek, entrenamenduko hiztegiaren % 31,3 osatzen dute. Aberastasun lexikoaren beste adierazle garrantzitsu bat hiztegitik kanpoko hitz ezberdinen (OOV, *out-of-vocabulary*) kopurua da. Hauek, garapen- eta proba-azpimultzoetan agertutako hitzak dira, zeinak entrenamendu-multzoan agertzen ez diren. Garapen-azpimultzoan hauen portzentaia % 8,55koa da; eta probaren kasuan, berriz, hiztegiaren % 8,89a osatzen dute. Entrenamendu-multzoan mezu asko egon arren, testu berriak prozesatzean hitz berriak agertzeko parada handia dela ikusten da. Laburtuz, corpusa lexikoki aberatsa da eta entrenamenduko hiztegitik at dauden hitzak azpimultzo bakoitzeko % 8 baino gehiago dira.

	entrenamendua	garapena	proba
Mezu kopurua	140.660 (% 80)	17.582 (% 10)	17.583 (% 10)
Klase-banaketa	<i>suicide</i> : % 38,9	<i>suicide</i> : % 38,9	<i>suicide</i> : % 38,9
	<i>non-suicide</i> : % 61,1	<i>non-suicide</i> : % 61,1	<i>non-suicide</i> : % 61,1
Hiztegiaren tam.	27.523	13.127	13.094
Hapax kopurua	8.609 (% 31,3)	4.930 (% 37,6)	4.913 (% 37,5)
OOV kopurua	-	1.122 (% 8,55)	1.163 (% 8,89)

6.2 Taula: Datu-sortaren partiketari buruzko datuak. OOV: *out-of-vocabulary*.

6.3 Artearen egoerako hurbilpena

Artearen egoerako hurbilpena Word2Vec hitz-embeddingek eta ikasketa sakonean oinarritutako sailkatzaileek osatzen dute. Kasu honetan, *Suicidal-BERT* lanean² erabilitako eredu eta entrenamendu prozesuak erreproduzitu dira, bertan erabiltzen baitira aurrekarien analisisian aurkeztutako hainbat eredu: Word2Vec hitz-embeddingak, ohiko neurona-sareak eta

²Kodea eskuragarri hemen: <https://github.com/gohjiayi/suicidal-text-detection/>

atentzio-mekanismoa erabiltzen duten transformerrak (eredu guztiei buruzko informazio gehiago [4.3.](#) atalean).

6.3.1 Errepresentazioa: hitz-embeddingak

Hurbilpen honetan mezuak errepresentatzeko Word2Vec hitz-embeddingak erabili dira. Horrela, mezu bakoitzaren errepresentazioa bere hitzen errepresentazio-bektoreek osatzen dute. Eredu honen aurre-entrenatutako bertsioak eskuragarri daude, milioika hitzen gainean entrenatuak izan direnak. Ohikoa da aurre-entrenatutako embeddingak erabiltzea hizkuntza orokorra lantzen duten atazetan, baita eskuragarri dagoen datuen bolumena embedding pertsonalizatuak lortzeko txikiegia denean ere. Kasu honetan, aldiz, lexikoki aberatsa den eta tamainaz esanguratsua den datu-sorta dugunez (entrenamendu-multzoko hiztegiaren tamaina 27.523 hitzekoa da), embeddingak datu-sortaren gainean entrenatzea erabaki da [\[62\]](#). Zehazki, Word2Vec ereduaren gaineko *fine-tuning* prozesu bat gauzatu da ereduak duen ezagutza proiektu honetan erabilitako hiztegiara egokitzeko.

Jarraian, *fine-tuning* prozesuaren ezaugarriak eta erabilitako hiperparametroak zerrendatzen dira:

- Erabilitako datuak: entrenamendurako azpimultzoa (prestaketa ostean lortutako datu-sortako mezuen % 80).
- Embeddingen dimentsio kopurua: 300.
- Entrenamenduko *epoch* edo iterazio kopurua: 5.
- Leihoa (embedding bakoitza entrenatzean kontuan hartuko diren aurreko eta ondorengo hitzen kopurua): 10.
- Hitzaren agerpen kopuru minimoa (entrenamenduan ez dira kontuan hartuko kopuru horretatik beherakoak dituzten hitzak): 2.

6.3.2 Sailkatzaileak: ikasketa sakonean oinarrituak

Sailkatzaileei dagokienean, artearen egoerako bost ezberdin landu dira, [4.3.3.3.](#) atalean azaldutakoak. Jarraian, eredu bakoitzaren entrenamendu prozesuari buruzko xehetasunak azaltzen dira.

6.3.2.1 Erregresio logistikoa

Nahiz eta erregresio logistikoa (*Logit*) ikasketa sakonean oinarritutako eredu bat ez izan, ohikoa da hizkuntzaren prozesamenduan sailkapenerako oinarri-lerro gisa erabiltzea. Kasu honetan ez da hiperparametro berezirik adierazi. Aipatu beharrekoa da, ordea, *Suicidal-BERT* laneko autoreek egindakoa errespetatuz, kasu honetan ez dela proiektu honetan zehaztutako ebaluazio-eskema jarraitu. Izan ere, haien hitzetan, oinarri-lerroan ez da garapen-azpimultzorik behar. Beraz, kasu honetan, aurretik finkatutako mezuen % 80 entrenamendurako erabili den arren, ereduaren ebaluaziorako gainontzeko % 20a erabili da, berez garapenerako eta probarako banandutako azpimultzoen elkarketa. Ereduaren sarrera gisa, prestatutako mezuetako hitzen Word2Vec bektoreen batezbestekoa erabili da.

Laburbilduz, hauek dira ereduaren ezaugarri nagusiak:

- Datuen partiketa: 80:20.
- Ereduaren sarrera: prestatutako mezueta hitzen Word2Vec errepresentazioen batezbestekoa.

6.3.2.2 Neurona-sare konboluzionala (CNN)

Neurona-sare konboluzionalen kasuan, *Suicidal-BERT* laneko autoreek diseinatutako arkitektura bat erabili da, geruza hauek dituena: embedding geruza bat (ataza honetarako entrenatutako Word2Vec embeddingekin osatua), lau konboluzio geruza, *max-pooling* geruza bat, diluzio (*dropout*) geruza bat eta guztiz konektatutako geruza bat.

Eredu honen entrenamendu eta proba prozesuan ebaluazio-eskeman finkatutako partiketa erabili da: prestaketa ostean lortutako datu-sortako mezuen % 80 eredia entrenatzeko erabili da, % 10 garapenerako eta beste % 10a probarako.

Jarraian, ereduaren hiperparametroak laburbiltzen dira:

- Datuen partiketa: 80:10:10.
- Ereduaren sarrera: prestatutako mezuak (entrenatutako Word2Vec embeddingak lortuko dira embedding geruzaren bitartez).
- *Batch*-tamaina: 32.
- Konboluzio geruza bakoitzeko filtro kopurua: 32.
- Konboluzio geruza bakoitzeko filtroen tamaina: 1. geruza: 5x5; 2. geruza: 6x6; 3. geruza: 7x7; 4. geruza: 8x8.
- Diluzio-tasa: 0,2.
- Optimizatzailea: Adam.
- Galera-funtzioa: Binary Cross Entropy with Logits Loss.
- Ikasketa-tasa: 0,00001.
- Aktibazio-funtzioa: ReLU.
- *Epoch* kopurua: 5.

6.3.2.3 LSTM

Kau honetan ere *Suicidal-BERT* lanean diseinatutako arkitektura erabili da LSTM eredia lortzeko, osagai hauek dituena: embedding geruza bat (ataza honetarako entrenatutako Word2Vec embeddingekin osatua), bi LSTM geruza, diluzio (*dropout*) geruza bat eta guztiz konektatutako geruza bat, sigmoide aktibazio-funtzioa duena.

Kasu honetan ere ebaluazio-eskeman finkatutako partiketa erabili da: prestaketa ostean lortutako datu-sortako mezuen % 80 eredia entrenatzeko erabili da, % 10 garapenerako eta beste % 10a probarako.

Hauek dira ereduaren ezaugarriak:

- Datuen partiketa: 80:10:10.
- Ereduaren sarrera: prestatutako mezuak (entrenatutako Word2Vec embeddingak lortuko dira embedding geruzaren bitartez).
- *Batch*-tamaina: 32.
- Ezkutuko-bektoreen tamaina: 128.
- Diluzio-tasa: 0,5.
- Optimizatzailea: Adam.
- Galera-funtzioa: Binary Cross Entropy with Logits Loss.
- Ikasketa-tasa: 0,00001.
- Aktibazio-funtzioa: Sigmoidea.
- *Epoch* kopurua: 1.

6.3.2.4 BERT

BERT transformerraren kasuan, BERT-base uncased eredu erabili da. Honek 12 geruza (ezkutuko-bektoreen tamaina 768 izanik) eta 12 atentzio-buru ditu (110 milioi parametro inguru). *Uncased* izateak hitz guztiak hizki xehera bihurtzen dituela esan nahi du.

BERT-base ereduaren tamainagatik eta hau entrenatzeko behar den datu eta baliabide konputazional kopuru erraldoiagatik, zerotik entrenatzeko aukera hasiera batetik baztertu da. Hori egin ordez, *fine-tuning* teknika erabili da eredu ataza honetara egokitzeko. Honela, entrenamendua zerotik hasi ordez, ereduak aurre-entrenamenduan ikasitako parametroen balioak erabili dira abiapuntu gisa, eta lan honetan erabili den datu-sortarekin egokitu. Hau egiteko bi aukera daude: eredu guztia egokitzea edota geruza batzuk izoztu eta gainerakoak soilik doitzea. Kasu honetan, lehen aukeraren alde egitea erabaki da.

Eredu honek lan honetan erabilitako neurona-sareekiko beste berezitasun bat du. Izan ere, jatorrizko BERT ereduak esaldi osoak erabiliz entrenatu dira. Beraz, ahalik eta emaitza onenak lortzeko, kasu honetan ez dira prestatutako mezuen hitz-embeddingak erabili sarrera gisa, mezu gordinak baizik, prestaketa egin aurrekoak. Mezu gordin horien gainean BERTen tokenizazioa erabili da, hitz bakoitzari eredu honek erabiltzen dituen identifikatzaileak esleitzeko.

Jarraian eredu honen ezaugarri eta hiperparametroak laburbiltzen dira:

- Datuen partiketa: 80:10:10.
- Ereduaren sarrera: mezu gordinak.
- *Batch*-tamaina: 6.
- Ikasketa-tasa: 0,00001.
- *Epoch* kopurua: 1.

6.3.2.5 ELECTRA

ELECTRAren kasuan ere ELECTRA-base uncased eredua erabili da, BERTen tamaina berekoa dena: 12 geruza (ezkutuko-bektoreen tamaina 768 izanik) eta 12 atentzio-buru ditu (110 milioi parametro inguru). Eredu honekin ere, *fine-tuning* prozesu bera jarraitu da, eredu osoa birdoitzuz. Sarrerari dagokionean, oraingoan ere mezu gordinak erabili dira, ELECTRAren tokenizatzailarekin tokenizatu direnak.

Erabilitako hiperparametroak eta ereduaren ezaugarriak BERTen erabilitako berak izan dira:

- Datuen partiketa: 80:10:10.
- Ereduaren sarrera: mezu gordinak.
- *Batch*-tamaina: 6.
- Ikasketa-tasa: 0,00001.
- *Epoch* kopurua: 1.

6.3.3 Lortutako emaitzak

Aurkeztutako 5 eredu gainbegiratuekin lortutako emaitzak 6.3. taulan ikus daitezke. Eredu guztiekin % 90etik gorako asmatze-tasak lortu dira, hau da, eredu guztiak gai izan dira probarako azpimultzoko 10 mezutatik 9 ondo sailkatzeko. Ohiko neurona-sareek eta atentzio-mekanismoa erabiltzen duten transformerrek lortutako emaitzak alderatuz gero, ikus daiteke lehenengoekin % 91,6 inguruko emaitzak lortu direla, eta bigarrenekin berriz, % 97,8 ingurukoak. Beraz, atentzioa erabiltzen duten ereduak ohiko neurona-sareen portaera nabarmenki hobetzen dutela esan daiteke, 6 puntu baino gehiagoko diferentziarekin. Zehazki, erabilitako metrika guztietan gailendu den eredu ELECTRA izan da, % 97,9ko asmatze-tasarekin.

Eredua	Asmatze-tasa	Doitasuna	Estaldura	F puntuazioa
Logit	90,4	87,4	88,1	87,7
CNN	91,5	90,4	87,5	88,9
LSTM	91,7	90,3	88,0	89,1
BERT	97,7	96,8	97,2	97,0
ELECTRA	97,9	97,0	97,7	97,4

6.3 Taula: Word2Vec hitz-embeddingekin eta artearen egoerako sailkatzailarekin lortutako emaitzak (%).

6.3.4 Eztabaida

Datu-sorta hau erabili duten aurrekariak lortutako emaitzak esperimazio honetan lortutakoekin alderatuz (informazio gehiago 4.1.2. atalean) hainbat ondorio atera daitezke. Batetik, balitekeela emaitza hobeak lortzea errepresentazio gisa Doc2Vec bektoreak erabilia Word2Vec hitz-embeddingak erabili ordez. Izan ere, *Suicide Prevention* lanean dokumentu-embedding horiek erabili dira mezuak errerepresentatzeko, eta ondoren erregresio logistikoa

sailkatzaile gisa erabiliz % 92,2ko asmatze-tasa lortu da, eredu horrekin lan honetan lortutakoa baino ia 2 puntu gehiago. Beraz, erregresio logistikoa oinarri-lerrotzat erabiltzen dela kontuan hartuta, baliteke errepresentazio modu horrekin emaitza hobeak lortzea lan honetan erabilitako neurona-sareekin. Hala ere, hau ezin da guztiz baieztatu, lan horrek datu-sortaren partiketa ezberdina erabiltzen duelako. Bestetik, *Suicidal-BERT* laneko ereduak erreproduzitututa jatorrizko emaitza berak lortu dira, ELECTRA transformerrarekin bi kasuetan % 97,9ko asmatze-tasa lortuz.

Hurbilpen honetan ohiko neurona-sareak (CNN eta LSTM) eta atentzio-mekanismoa erabiltzen duten transformerrak (BERT eta ELECTRA) erabili dira sailkatzaile gisa. Bi eredu mota hauekin lortutako emaitzak alderatuz ikusi da transformerrek ohiko neurona-sareen portaera nabarmenki hobetzen dutela, lortutako asmatze-tasen diferentzia 6 puntu ingurukoa izanik. Honek badu zentzua, hasteko, proiektu honetan erabilitako transformerrak milioika hitzeko corpusen gainean aurre-entrenatuak izan direlako, CNN eta LSTM neurona-sareak zerotik eta soilik lan honetan erabili den datu-sortaren gainean entrenatu diren bitartean. Gainera, bi eredu mota hauen tamainan diferentzia esanguratsuak daude, atentzioa erabiltzen duten ereduak parametro askoz ere gehiago baitituzte.

Lan honetan ELECTRA transformerrarekin lortutako emaitzetan sakontzeko asmoz, ereduaren nahasmen-matrizea eraiki da, 6.4. taulan ikus daitekeena. Negatibo-falstu (NF) erroreek suizidio-ideiagintzaren zantzuak dituzten eta sistemak hala identifikatu ez dituen mezuak errepresentatzen dituzte. Positibo-falstu (PF) erroreak aldiz aurkako kasuetan gertatzen dira, benetan suizidioarekin zerikusirik ez duten mezuak sistemak gai horrekin erlazioa duten mezu gisa sailkatzen dituenean. Ataza honetan NF errorea minimizatzea komeni da, garrantzitsuagoa baita suizidio-zantzuak dituzten mezu guztiak identifikatzea honen arrastorik erakusten ez duen mezu bat *suicide* gisa etiketatzea baino. ELECTRArekin lortutako emaitzen kasuan, NF errore kopurua 158 mezuak izan da, PF erroreena handiagoa izan den bitartean, 204 mezuak, hain zuzen ere. Hau 6.3. taulan ere ikusten da, ELECTRA ereduaren estaldura % 97,7koa baita, doitasuna baxuagoa den bitartean, % 97koa.

		Iragarritako etiketa		
		P'	N'	guztira
Etiketa errealak	P	6.681 EP	158 NF	6.839
	N	204 PF	10.540 EN	10.744
guztira		6.885	10.698	

6.4 Taula: ELECTRA transformerrarekin lortutako emaitzen nahasmen-matrizea.

NF eta PF erroreak ilustratzeko asmoz, jarraian, adibide bana aurkezten da kasu bakoitzerako:

NF: *Living 18 years is enough right? I mean, 18 years sounds like a lot. 18 years of anything sounds like too much. 18 years of swimming, studying, eating, all sound too much. 18 years of life is way too much*

PF: *why everyone use me ? hi, don't get me wrong i love helping people but when i help so much they depend on me more i feel like getting suffocated and i can't say no either because i have some fear.*

NF errorearen zergatia aztertuz, baliteke sistema ez ohartzea mezuak suizidioarekin zerikusia duela printzipioz suizidio-ideiagintzarekin erlaziorik ez duten ekintza asko aipatzen direlako: igeri egitea, ikastea eta jatea. Mezu honen giltza, ordea, “*too much*” eta “*enough*” hitzetan dago, testuinguru honetan negatibotzat jo daitezkeenak. Hau da, mezua idatzi duen pertsonak uste du 18 urtez bizitzea nahikoa eta gehiegi dela. Baliteke ereduak errorea egin izana fokua aipatzen diren ekintzetan jarri duelako, eta ez negatibotzat har daitezkeen zatietan, 18 urte bizitzea gehiegizkoa dela diotenak.

PF erroreari dagokionean, baliteke sistemak mezua suizidio-ideiagintzarekin lotuta dagoela erabakitzea bertan konnotazio negatiboa duten hitzak agertzen direlako, “*suffocated*” (itota) eta “*fear*” (beldurra), testuinguruaren arabera ideia gintza suizidarekin erlazionatu daitezkeenak.

Errore hauen analitika ondorioztatu daiteke sistema gai dela suizidio-ideiagintzarekin erlazionatutako hiztegi eta esamoldeak detektatzeko eta honekin zerikusirik ez dutenen artean bereizteko, baina batzuetan ez duela testuinguruan behar beste erreparatzen eta honek akatsak sorrarazten dituela. Hala ere, esperimentazioak erakutsi du artearen egoerako sistemek emaitza oso onak ematen dituztela ataza honetan, eta atentzio-mekanismoa erabiltzen duten ereduak, transformerrek, ohiko neurona-sareen portaera nabarmenki hobetzen dutela, % 97,9ko asmatze-tasa lortzera iritsiz.

6.4 Hurbilpen oinarritzkoagoa

Artearen egoerakoaren aldean oinarritzkoagoa den hurbilpena LDA topiko-ereduetatik abiatuta lortutako errepresentazioak eta ikasketa automatikoan oinarritutako sailkatzaileak osatzen dute. Hurrengo ataletan hurbilpen hau osatzeko egindako esperimentazioa eta lortutako emaitzak azaldu eta eztabaidatzen dira.

6.4.1 Errepresentazioa: topiko-ereduak

Kasu honetan mezuak errepresentatzeko topiko-ereduak erabili dira, zehazki, LDA (Latent Dirichlet Allocation) teknika. Hau erabiliz, mezuak topikoen gaineko probabilitate-banaketa gisa laburbiltzen dira bektoreen bitartez. Bektore hauek mezuen dimentsionalitate baxuko errepresentazio gisa erabiltzea erabaki da proiektu honetan.

Errepresentazio hau lortzeko, Word2Vec hitz-embeddingen *fine-tuning* prozesuan erabilitako ezaugarriak mantentzeko saiakera egin da, bien arteko alderaketa baldintza bertsuetan egiteko asmoz. Eredua osatzeko hitz-embeddingen entrenamenduan erabilitako datu-sortaren azpimultzo bera erabili da, prestaketa prozesu osoaren (aurreprozesaketa eta garbiketaren) ostean lortutako mezuen % 80ak osatua. Kasu honetan ere, agerpen kopurua 2 baino txikiagoa duten hitzak alde batera utzi dira.

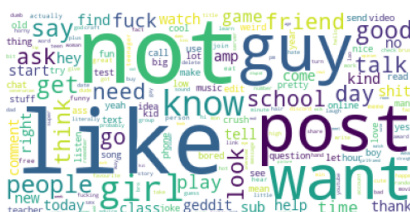
6. MEZUEN SAILKAPENA

Errepresentazioen tamainari, hau da, ereduaren topiko kopuruari dagokionean, hainbat balio ezberdin erabili dira hau aldatzeak sailkapenean izan dezakeen eragina ikusteko asmoz. Guztira 9 eredu ezberdin lortu dira, topiko kopuru (m) hauekin: 2, 5, 10, 50, 100, 150, 200, 250 eta 300.

Laburbilduz, ezaugarri hauek erabili dira LDA eredutik abiatutako errepresentazioak lortzeko:

- Erabilitako datuak: entrenamendurako azpimultzoa (prestaketa ostean lortutako datu-sortako mezuen % 80).
- Topiko kopurua: 2, 5, 10, 50, 150, 200, 250, 300.
- Hitzaren agerpen kopuru minimoa (ez dira kontuan hartuko kopuru horretatik beherakoak dituzten hitzak): 2.

Adibide gisa, 6.3. irudian $m = 2$ topikoko ereduarekin lortutako topiko bakoitzaren hitz-hodeiak aurki daitezke. Bertan, bi topikoei dagozkien hitz-banaketak adierazten dira, non hitzaren tamaina honek topikoan duen garrantziarekiko proportzionala den. Hitz-hodei hauek aztertuz, ikus daiteke badaudela bietan agertzen diren hitzak, “not”, “want”, “like” eta “know” esaterako (euskaraz ez, nahi, atsegin eta jakin, hurrenez hurren). Hala ere, bietan erabiltzen den lexikoan ezberdintasunak antzematen dira. 1. topikoan “post”, “guy”, “school”, “girl” (post, mutil, eskola, neska) eta antzerako hitzek dute presentziarik handiena. Lexiko hau nerabeek egunerokotasunean erabiltzen dutenarekin erlazionatu daiteke. 2. topikoan aldiz, “feel”, “life”, “die”, “think” eta “time” (sentitu, bizitza, hil, pentsatu, denbora) dira nabarmentzen direnak. Hauek, bizitzarekin eta heriotzarekin lotutakoak dira. Badirudi bigarren topikoan suizidio-ideiagintzarekin lotutako lexikoa ageri dela.



(a) 1. topikoaren hitz-hodeia.

(b) 2. topikoaren hitz hodeia.

6.3 Irudia: $m = 2$ topikoko ereduarekin lortutako topikoen hitz-hodeiak.

6.4.2 Sailkatzaileak: ikasketa automatikoan oinarrituak

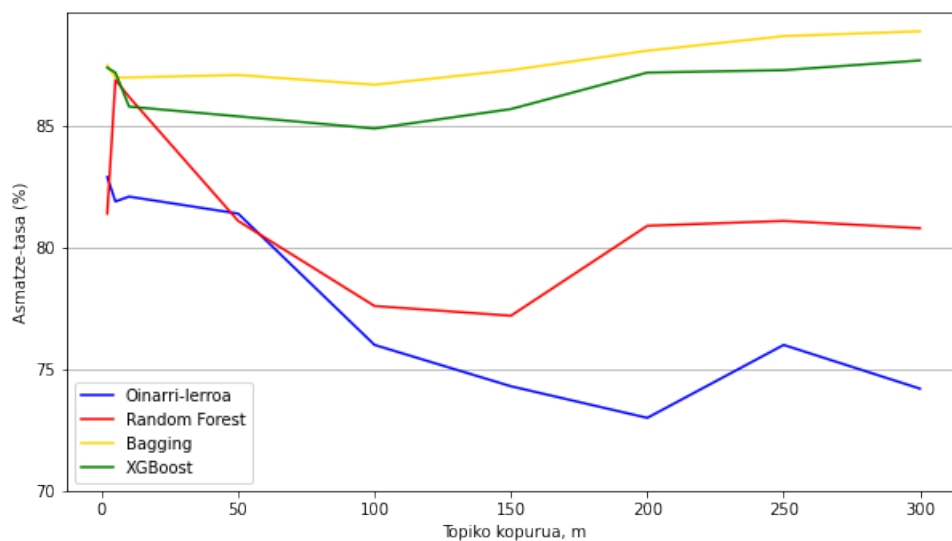
Hurbilpen honetan ikasketa automatikoan oinarritutako hiru meta-sailkatzaile erabili dira, 4.3.3.2. atalean azaldu direnak: Random Forest; bagging teknika erabiliz lortutakoa, oinarri gisa SVC sailkatzailea duena; eta XGBoost. Hauek Pythoneko sklearn liburutegia erabiliz inplementatu dira, hiperparametro gisa liburutegi honen besterik ezekoak erabiliz. Hiru

eredu horien oinarri-lerro gisa lan honetan proposatutakoa erabili da, LDA errepresentazioetan gailentzen den topikoa erabiltzen duena erabakia hartzeko (informazio gehiago 4.3.3.1. atalean).

Sailkatzaile hauen guztien kasuan ez da ebaluazio-eskemako datu partiketa zehatz-mehatz erabili. Izan ere, eredu hauek ez dute garapenerako azpimultzorik behar, ez baitira beren parametroak doitzen. Hori dela-eta, eredu hauen entrenamendurako aurretik finkatutako datuen % 80a erabili da, eta ebaluaziorako berriz, probarako hautatutako mezuak, sortaren % 10a, bi hurbilpenen emaitzak ahalik eta konparagarrienak izan daitezten. Garapenerako azpimultzoa alde batera utzi da.

6.4.3 Lortutako emaitzak

LDA erabiliz lortutako topiko kopuru ezberdineko errepresentazioen gainean entrenatu dira aipatutako lau sailkatzaileak. Kasu bakoitzean lortutako emaitzak A eranskinean aurki daitezke. Emaitza horietatik abiatuta 6.4. irudiko grafikoa sortu da, non landutako lau sailkatzaileek lortutako asmatze-tasak topiko kopuruarekiko izandako bilakaera azaltzen den.



6.4 Irudia: Sailkatzaile klasikoekin asmatze-tasak (%) izandako bilakaera topiko kopuruaren arabera.

Oinarri-lerroaren kasuan ikusi daiteke topiko kopurua handitu ahala asmatze-tasak behera egiteko tendentzia duela. Honen arrazoia oinarri-lerroan mezu bakoitzeko kontuan hartzen den topikoa bakarra dela da. Topiko kopurua geroz eta handiagoa izan, mezuaren edukia topiko gehiagotan sakabanatzea gerta daiteke, eta ondorioz topiko bakarrarekin hau sailkatzeko informazio nahikorik ez izatea. Hala ere, emaitzak ikusita oinarri-lerro funtzioa betetzen duela esan daiteke. Aipatzekoa da nahiz eta joera beherakorra erakutsi, $m = 250$ kasuan asmatze-tasak gorakada bat izan duela.

Random Forest sailkatzailearen kasuan gauza bitxia gertatzen da. Izan ere, asmatze-tasak gorakada izan du errepresentazioa 2 topikotik 5 topikora zabaltzean. Hortik aurrera, ordea, beherako joera hartzen du, 150 topikorekin minimora heldu arte. Hortik aurrera, asmatze-tasak hobera egiten du, baina 5 topikorekin lortutako asmatze-tasatik urrun gelditzen

da. Honen arrazoia zein den jakiteko Random Forest sailkatzaileak topiko kopuruaren aldakortasunarekiko duen sentiberatasuna analizatu beharko litzateke, proiektu honen irismenetik kanpo gelditzen dena.

Bagging teknika erabiliz osatutako meta-sailkatzailearen kasuan, asmatze-tasak tentsioa positiboa erakusten du topiko kopuruaren hazkundearekiko. Hala ere, hobekuntza hori oso txikia da topiko kopuruaren aldaketarekin alderatuta.

XGBoosten kasuan antzekoa gertatzen da. Asmatze-tasak beherako joera arina erakusten du topiko kopurua 100era iristen den arte. Hortik aurrera emaitzak hobetu egiten dira, baina 2 topikorekin eta 300ekin lortutako emaitzak oso antzekoak dira.

Orokorrean, emaitzarik onenak bagging teknikarekin lortu dira ia topiko kopuru guztietan. Hala ere, eta oinarri-lerroa kontuan izan gabe, 200 topikotik aurrera asmatze-tasaren hobekuntza gelditu egiten dela esan daiteke, eta ez dela % 90era iristen kasu bakar batean ere. Gainera, azpimarratzekoa da estaldura baliorik altuenak Random Forest sailkatzailearekin lortu direla, eta % 90era iritsi direla.

Asmatze-tasaren eta ereduaren ebaluaziorako erabilitako gainontzeko hiru neurrien bilakaerari buruzko datu zehatzagoak izateko, 6.5. taulan bildu dira $m = 5, 150, 300$ topiko kopuruekin lortutako emaitzak. Balio hauek aukeratu dira egindako proben bilakaera islatzen dutelako, aztertutako artean balio txikietako bat, erdibideko bat eta maximoa baitira. Topiko kopuru guztietan lortutako emaitzak, esan bezala, A eranskinen aurki daitezke.

m	Sailkatzailea	Asmatze-tasa	Doitasuna	Estaldura	F puntuazioa
5	Oinarri-lerroa	81,9	72,5	85,8	78,6
	Random Forest	86,9	82,2	84,8	83,5
	Bagging	87,0	81,8	85,5	83,6
	XGBoost	87,2	81,9	86,2	84,0
150	Oinarri-lerroa	74,3	67,1	66,6	66,9
	Random Forest	77,2	64,8	90,4	75,5
	Bagging	87,3	85,3	81,3	83,3
	XGBoost	85,7	81,8	81,2	81,5
300	Oinarri-lerroa	74,2	68,5	62,4	65,3
	Random Forest	80,8	69,4	90,5	78,6
	Bagging	88,9	87,4	83,4	85,4
	XGBoost	87,7	83,8	84,7	84,2

6.5 Taula: $m = 5, 150, 300$ topikoko LDA errepresentazioen gainean sailkatzaile klasikoekin lortutako emaitzak (%).

Topiko kopurua handitu ahala asmatze-tasak bilakaera handirik ez duela ikusten da 6.5. taulan. Bagging teknika erabiliz eraikitako meta-sailkatzailearen kasuan 1,9 puntuko hazkundera izan du topiko kopurua 5etik 300era handitzean. XGBoosten kasuan tarte berean asmatze-tasa 0,5 hazi da soilik, eta Random Foresten behera egin du nabarmenki, 6,1 puntu gutxiago lortuz. Emaitza onenak lortu dituen konbinazioa bagging motako meta-sailkatzailea eta $m = 300$ topikoko errepresentazioa izan da, % 88,9ko asmatze-tasarekin.

6.4.4 Eztabaida

Hurbilpen honen emaitzak artearen egoerako sistemekin lortutakoekin alderatuta ezberdintasunak ikus daitezke. Artearen egoerako hurbilpenean asmatze-tasarik altuena lortu duen eredua ELECTRA transformerra izan da, % 97,9koa izanik. Honetan berriz, gailendu den eredua bagging motako meta-sailkatzaileak eta $m = 300$ topikoko errepresentazioak osatua izan da, % 88,9ko asmatze-tasarekin. Bien arteko diferentzia 9 puntukoa da. Ohiko neurona-sareekin lortutako emaitzekin alderatzean, aldiz, diferentzia 2,8 puntura jaisten da. Beraz, ikasketa automatikoan oinarritutako sailkatzaileekin eta topiko-ereduekin artearen egoerako ohiko neurona-sareen errendimendura hurbiltzea lortu da, ez ordea transformerren emaitzetara. Honek erakutsi du posible dela topiko-ereduak erabiliz mezuak numerikoki errepresentatzea, eta baita topikoak erabiliz suizidio-ideiagintzaren zantzuak detektatzea ere.

Oinarritzkoagoa den hurbilpen honekin lortutako emaitzetan fokua jarrita, ikusi da topiko kopurua handitzen joan ahala landutako sailkatzaileak azkar saturatzen direla, hau da, nahiz eta topiko gehiagoz osatutako errepresentazioak erabili, emaitzak ez direla hobetzen. Are gehiago, Random Forest ereduaren kasuan, asmatze-tasa baxuagoa lortu da 300 topikoko errepresentazioarekin (% 80,8) 5 topikorekin baino (% 86,9). Baliteke honen arrazoia datu-sorta honetako mezuen edukia laburbiltzeko topiko gutxi nahikoa direla izatea, eta kopuru hau handitzeak emaitzetan onurarik ez islatzea.

Asmatze-tasarik altueneko ereduaren portaeran sakontzeko asmoz ($m = 300$ eta bagging meta-sailkatzailea), honen nahasmen-matrizea eraiki da, 6.6. taulan aurki daitekeena. Kasu honetan negatibo-faltsuen kopurua positibo-faltsuena baino nabarmenki handiagoa da. Aurretik esan bezala, ataza honetan ahalik eta negatibo-faltzu gutxien egotea bilatzen da, suizidio-ideiagintza eztabaidatzen duten mezu guztiak identifikatzea baita lehentasuna. Hau nabaria da 6.5. taulan ere bai, eredu honen estaldura doitasuna baino baxuagoa baita, % 83,4koa.

		Iragarritako etiketa		
		P'	N'	guztira
Etiketa errealak	P	5.702 EP	1.137 NF	6.839
	N	819 PF	9.925 EN	10.744
guztira		6.521	11.062	

6.6 Taula: $m = 300$ topikoko LDA errepresentazioa eta bagging teknika erabiliz eraikitako meta-sailkatzailearekin lortutako emaitzen nahasmen-matrizea.

Errore hauei buruzko informazio gehiago izateko, jarraian NF eta PF errorearen adibide bana aurkezten da:

NF: *please help me I cant breath I feel like my chest is caving in I cant stop crying why wont anyone help me??*

PF: *what if u sneeze while they put the stick up ur nose for the corona test? ive always had it down my throat but im about to get it up the nose and im scared cause what if i sneeze and it launches me forward and the stick gets impaled inside the back wall of my nose where it just so happens that the weakest part of the skull is lokated and i fucking die? worst case scenario but still i cant imagine not sneezing having something put up ur nose*

Adibide gisa aurkeztutako NF errorea gertatu den mezuan, erabiltzaileak laguntza eskaera bat egiten du, arnasteko ezintasuna duela eta etengabe negarrez ari dela argudiatuz. Sistemak, ordea, ez du mezu honetan suizidio-ideiagintzaren zantzurik aurkitu. Honen arrazoia mezuan esplizituki heriotza edota suizidioarekin erlazonatutako lexikorik ez egotea izan daiteke. PF errorearen kasuan, aldiz, aurkakoa gertatu da. Izan ere, mezu honetan “die” hitza agertzen da, baina ez suizidio-ideiagintzarekin erlazonatuta. Analisi honetatik ondoriozta daiteke balitekeela eredu honek heriotza eta suizidioari buruzko lexiko espezifikoan fokua jartzea, eta hauei inplizituki egiten zaizkien erreferentziaz ez ohartzea.

ELECTRA transformerraren errore-analisan (ikus 6.3.4. atala) aurkeztutako NF errorea adibidea kasu honetan ere gaizki sailkatua izan da, baina hurbilpen honetan adierazitako NF errorea ondo sailkatu du ELECTRAk. PF errorearen kasuan, ELECTRAk gaizki sailkatutako adibidea ondo sailkatua izan da bagging meta-sailkatzailearekin, eta baggingen kasuan jarritako errorearen adibidea ondo sailkatu da ELECTRArekin. Hurbilpen bakoitzean emaitzarik onenak eman dituzten ereduak egindako erroreak alderatuz, ELECTRA ereduak egindako errorearen % 39,8a bagging ereduak zuzen sailkatu du eta meta-sailkatzaile horrek egindako errorearen % 88,9a berriz, ondo sailkatu du ELECTRAk. Zentzuzkoa da ELECTRAk baggingen errore gehiago ondo sailkatu izana, asmatze-tasa altuagoa duelako. Hala ere, azpimarratzekoa da aurkako kasuan gertatutakoa, hau da, bagging teknika erabiliz lortutako sailkatzaileak transformerrak egindako errorearen ia % 40a era egokian sailkatzen dituela. Honek erakusten du eredu bakoitzak mezuen alderdi ezberdinetan jartzen duela arreta, eta balitekeela biak nolabait konbinatuta emaitzak hobetzea.

Laburbilduz, esperimentu hauekin ikusi da topiko-ereduek mezuak errepresentatzeko ahalmena dutela, eta errepresentazio hauek baliagarriak direla mezuetan suizidio-ideiagintza detektatzeko. Ikasketa automatikoa oinarritutako sailkatzaileekin emaitza onak lortu diren arren, % 88,9ko asmatze-tasara iritsiz, ikusi da sailkatzaile hauek topiko kopurua handitzearekin saturatzen direla, eta topiko gehiago erabiltzeak ez duela lortutako emaitzen kalitatea apenas hobetzen.

6.5 Sailkapenaren laburpena

Egindako esperimentazioarekin ikusi da artearen egoerako sistemek, hitz-embedding eta ikasketa sakonean oinarritutako sailkatzaileek, emaitza oso onak lortzen dituztela mezuen gaineko sailkapen bitarrean. Landutako ereduaren artean transformerrak nabarmendu dira bereziki, ELECTRArekin % 97,9ko asmatze-tasa lortu baita.

Bestetik, LDA topiko-ereduak eta ikasketa automatikoa oinarritutako sailkatzaileek osatutako hurbilpenarekin, topiko-ereduak mezuak errepresentatzeko balio dutela frogatu

da, eta errepresentazio hauek baliagarriak direla mezuetan suizidio-ideiagintza detektatzeko. Kasu honetan emaitzarik onenak lortu dituen eredia 300 topikoko errepresentazioak eta bagging teknika erabiliz lortutako meta-sailkatzaileak osatua izan da, % 88,9ko asmatzetasarekin. Aipatu beharrekoa da sailkatzaile klasikoak topiko kopurua handitzearekin saturatu egin direla, eta topiko gehiago erabiltzeak ez duela lortutako emaitzen kalitatea era esanguratsuan hobetzea lortu.

Gainera, erroreen analisitik ondorioztatu da balitekeela bi hurbilpenak nolabait konbinatuta batak bestearen erroreak zuzentzen laguntzea, lortutako emaitzak hobetuz.

Ondorioak eta etorkizuneko lana

Proiektu honetan sare sozialen testuinguruan suizidio-ideiagintzaren detekzioarekin erlazioa duten bi ataza landu dira: mezuen analisi linguistikoa eta sailkapen bitarra. Horretarako lehendabizi aurrekarien azterketa egin da eta hau kontuan izanda bi atazak ebatzi dira hizkuntzaren prozesamenduko teknika ezberdinak erabiliz. Lan osoan zehar Reddit sare-soziala iturri duen, ingelesez idatzitako mezuak biltzen dituen eta publikoki eskuragarri dagoen *Suicide and Depression Detection* datu-sorta erabili da.

Kapitulu honetan, egindako lanetik ateratako ondorioak aurkezten dira: zientifiko-teknikoak, proiektuaren kudeaketarekin erlazionatuak eta pertsonalak. Gainera, etorkizunerako norabideak proposatzen dira.

7.1 Ondorio zientifiko-teknikoak

Mezuen analisi linguistikoari dagokionean, proiektu honetan beste egile batzuek aztertutako lau ezaugarri landu dira aukeratutako datu-sortaren gainean: mezuen luzera, galderamarkak, lehen pertsona singularreko izenordainak eta aditz-denborak. Analisi honetan ikusi da datu-sorta honen kasuan betetzen dela suizidio-ideiagintzarekin erlazionatutako mezuak luzeagoak izatea eta lehen pertsona singularreko izenordain gehiago erabiltzea. Galderamarkekin eta aditz-denborekin, ordea, ez da esperotakoa bete. Izan ere, bi klaseetako mezuak galdera-marken erabilera bertsua egiten dute, eta suizidio-ideiagintzarekin zerikusirik ez duten mezuetan ikur hauen kopurua apur bat handiagoa da. Aditz-denboren erabilerari dagokionean, egindako analisitik ondorioztatu da ez dagoela bi klaseen arteko ezberdintasun esanguratsurik. Beraz, proiektu honetan mahaigaineratutako lehen ikerketa-galderari (IG1) erantzunez, *Suicide and Depression Detection* datu-sortan aurrekarietan identifikatutako ezaugarrietako batzuk bereizten dira, ez denak, ordea.

Hala ere, kontuan izan behar da mezuak bi *subreddit* ezberdinetatik eskuratuak izan direla eta bakoitzaren jatorriaren arabera esleitu zaiela 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.

Proiektu honetan, 4.2. atalean azaldu bezala, analisi linguistikorako ideiagintza suizidaren detekzioaren arloan erabiltzen den LIWC softwarearen [22] ordez, Python programazio lengoaiako hainbat oinarrizko funtzio eta ikasketa sakonean oinarritutako UDPipe analizatzailea [41] erabiltzea erabaki da. Honen arrazoia, LIWC hiztegiaren oinarritutako tresna dela izan da, eta ondorioz hainbat murriztapen dituela, hitz bakoitzaren testuinguru kontuan hartzen ez duela, esaterako. Lan hau hizkuntzaren prozesamenduaren testuinguruan kokatzen denez, ikasketa sakonean oinarritutako analizatzaileen alde egitea erabaki da, hitzen kategoria gramatikala eta ezaugarri linguistikoak jaso eta hizkuntzaren aldetik analisi sakonagoa egiteko aukera eskaintzen dutelako.

Sailkapenaren kasuan, ikusi da artearen egoerako sistemek emaitza oso onak ematen dituztela, ia % 98ko asmatze-tasara iritsiz ELECTRA transformerrarekin. Aztertutako ereduen artean ikusi da atentzio-mekanismoa erabiltzen duten transformerrek ohiko neurona-sareek lortutako emaitzak hobetzen dituztela. Proiektu honetan proposatutako oinarrizko hurbilpenarekin, aldiz, ia % 89ko asmatze-tasa lortu da. Honek erakutsi du topiko-ereduek badutela mezuak errepresentatzeko ahalmena, eta posible dela topikoak erabiliz suizidio-ideiagintzaren zantzuak detektatzea. Gainera, ikusi da ikasketa automatikoko sailkatzaileak azkar saturatzen direla topiko kopurua handitzen joan ahala.

Bi hurbilpenetako bakoitzean emaitza onenak eman dituen eredu kontuan hartuta, Word2Vec (300 dimentsio) - ELECTRA eta LDA (300 topiko) - bagging, asmatze-tasen arteko diferentzia 9 puntukoa da. Hala ere, ohiko neurona-sareekin (CNN eta LSTM) lortutako emaitzak oinarrizko hurbilpenarekin lortutakoekin alderatuta, diferentzia hori nabarmenki jaisten da 2,8 puntura. Beraz, lan honen bigarren ikerketa galderari (IG2) erantzunez, esan daiteke oinarrizkoagoa den hurbilpen bat erabiliz gerturatu gaitezkeela artearen egoerako ereduarekin lortutako emaitzetara. Gainera, ikusi da hurbilpen bakoitzean eraikitako ereduak mezuen alderdi ezberdinetan jartzen dutela arreta, eta posible litzatekeela bata bestearen osagarri gisa erabiltzea.

Hurbilpen ezberdinek lortutako emaitzetatik haratago joanda, badaude azpimarratzeko beste hainbat alderdi. Artearen egoerako sistemek, neurona-sare sakonetan oinarrituak, kutxa-beltz gisa funtzionatzen dute eta ez dute argi uzten zertan oinarritzen diren erabaki horiek hartzeko. Hauen aldean, LDA topiko-ereduak eskaintutako errepresentazioak askoz ere interpretagarriagoak dira, bektoreko posizio bakoitza topiko batekin erlazionatua baitago, mezu batek topiko bakoitzarekin duen erlazioaren proportzioa adieraziz. Gainera, erraza da topiko bakoitzaren hitz-banaketa era grafikoan aurkeztea hitz-hodeiak erabiliz, esaterako. Horrez gain, ikasketa sakonean oinarritutako sistemak entrenatzeak kostu konputazional handiagoa dakar LDA ereduak eta sailkatzaile klasikoak erabiltzeak baino. Beraz, emaitzak ikusita, asmatze-tasa altuak lortzea eta ereduen interpretagarritasunaren alde egitea balantza batean jarri beharreko alderdiak direla esan daiteke.

Laburbilduz, Gradu Amaierako Lan honen bitartez bi ekarpen zientifiko egin dira. Alde batetik, lehen pausoak eman dira *Suicide and Depression Detection* datu-sortaren gaineko analisi linguistikoan, hainbat ondorio esanguratsu lortuz. Bestetik, frogatu da topiko-ereduak baliagarriak direla mezuak errepresentatzeko eta sailkapenean emaitza onargarriak ematen dituztela.

Proiektu honetan egindako lanari esker artikulua zientifiko bat ekoitzi da, V. IkerGazte euskal ikertzaile gazteen kongresuan argitaratua eta aurkeztua izan dena:

Sara Gracia, Maite Oronoz, and Alicia Pérez. *Ideia gintza suizidaren identifikazioa sare sozialetan*. V. Ikerlazio Nazioarteko Ikerketa Euskaraz Kongresuko artikulu-bilduma. Ingeniaritza eta Arkitektura, 2023, 123–130 orrialdeak. ISBN/ISSN: 978-84-8438-865-4

7.2 Proiektuaren kudeaketarekin erlazionatutako ondorioak

Gradu Amaierako Lanaren irismena handia da, 2. kapituluaren adierazi den bezala. Honek ezer baino lehen plangintza on bat diseinatzea eskatzen du. Hala ere, ohikoa izaten da proposamenarekiko desbiderapenak sortzea eta honek ezinbestekoa egiten du proiektuaren garapena bizi-ziklo osoan zehar kudeatzea.

Orokorrean lan honetan definitutako plangintza bete da, 7.1. taulan ikus daitekeenez. Guztira estimatutakoa baino 15 ordu gehiago behar izan dira lana burutzeko, aurreikusitako denboraren % 5 gehiago, tamaina honetako proiektuetan onargarria dena. Desbiderapenik esanguratsuen memoriaren idazketak sortu du. Izan ere, dokumentu luzea da eta proiektuari buruzko informazio guztia jaso behar du. Zailtasunak izan ditut honen egitura definitzerako garaian, atalen antolamenduan eta atal bakoitzean eman beharreko azalpenak zehaztean. Idazketa prozesuan aurrera egin ahala dokumentuaren egitura finkatzen joan da beste lan batzuk irakurtzeari eta zuzendarien iradokizunei esker. Denboraren inguruko desbiderapenak aintzat hartu ziren proiektuaren arrisku-analisia gauzatu zenean, eta nahiz eta badaezpadako ordu-koltxoi hori erabili den, denbora gehiago behar izan da dokumentazioa idazteko. Hala ere, desbiderapen honek ez du proiektuaren bideragarritasuna kolokan jarri.

Adarra	Estimazioa	Errealak	Desbiderapena
Proiektuaren kudeaketa	40	35	-5 (% 12,5)
Arloaren egoeraren azterketa	70	70	-
Esperimentuen diseinua eta inplementazioa	95	90	-5 (% 5,3)
Dokumentazioa	95	120	+25 (% 26,3)
GUZTIRA	300	315	+15 (% 5)

7.1 Taula: LDEn definitutako adar bakoitzerako egindako estimazioak eta izandako desbiderapenak ordutan.

Proiektuaren helburu zientifiko-teknikoei dagokienean gustatuko litzaidake analisi linguistikoa gehiago sakontzeko aukera izatea, ezaugarri ezberdin gehiago landuz. Hala ere, lau ezaugarri aztertu dira eta emaitza esanguratsuak lortu dira. Gehiago ez sakontzearen arrazoia denbora falta izan da, aurrekariak identifikatutako ezaugarri asko analizatzaile linguistiko bidez aztertzeak abstrakzio-maila altuagoa eskatzen duelako, eta esperimentazio prozesu luzeak gauzatzea. Tankera honetako ezaugarriak aztertzea etorkizunerako norabide gisa proposatzea erabaki da, hurrengo atalean azaltzen den bezala.

Hau guztia jakinda, azken hausnarketa bezala esango nuke ezinbestekoa dela lehen momentutik proiektuaren irismena finkatzea, bai helburu aldetik eta baita denbora aldetik ere, eta horren arabera plangintza zehatz bat egitea. Gainera, memoriaren idazketak denbora asko eskatzen duela nabarmenduko nuke, eta plangintza diseinatzerako orduan kontuan izan beharreko zerbait dela.

7.3 Etorkizunerako lana

Egindako lanarekin ikusi da suizidio-ideiagintza eta sare sozialen ikerketa-esparruan bada-goela zer egina, eta badagoela hobekuntzarako espazioa. Jarraian, etorkizunean lan honi jarraipena emateko hainbat proposamen egiten dira.

Analisi linguistikoari dagokionez, proiektu honetan hasitako analisiak ateak ireki dizkio abstrakzio-maila altuagoko ezaugarriak aztertzeari. Izan ere, aurrekariak suizidio-zantzuak erakusten dituzten mezuetan ezeztapena presentego 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 analisisan sakontzea eskatzen dute, kategoria gramatikal edota dependentzia erlazioetatik haratago joan eta hitzen semantika aztertzea.

Sailkapenean, ikusi da topiko-ereduak erabiliz lortutako errepresentazioak lagungarriak izan daitezkeela sare sozialetako mezuetan suizidio-ideiagintza identifikatzeko. Ildo horretatik jarraituta, sailkatzaile klasikoetatik haratago joan eta neurona-sareak erabil litezke emaitzak hobetzen diren ikusteko. Beste aukera bat errepresentazio hau beste ezaugarri iragarle zein errepresentazioekin konbinatzea litzateke. Palani et al. autoreen lanean [63], esaterako, BERT transformerrarekin lortutako testuinguru-embeddingak (*contextual-embeddings*) eta LDA topiko-ereduarekin lortutako errepresentazioak konbinatzen dira txioen gaineko sentimendu analisia egiteko. Lan honetan ondorioztatzen da bi eredu hauen (testuinguru-embeddigen eta LDA topiko-ereduaren) konbinazioak BERT soilik erabilia lortutako emaitzak hobetzen dituela. Honek proiektu honetan ateratako ondorio bat indartzen du, hurbilpen oinarritzkoagoan lortutako eredu ELECTRA transformerraren osagarria izan daitekeela eta honen erroreak zuzentzen lagun dezakeela.

Beste proposamen bat proiektu honetan landutako bi atazak uztartzea litzateke, hau da, sailkapenerako iragarle gisa mezuen ezaugarri linguistikoak erabiltzea. Ikusi da topiko kopuru txikiak erabiliz (5 topikokoko errepresentazioarekin, adibidez) eta ikasketa automatikoan oinarritutako sailkatzaileekin % 80tik gorako asmatze-tasak lortu direla. Posible litzateke ezaugarri linguistiko gutxi batzuk erabiliz antzeko emaitzak lortzea? Eredu sinpleak lirarteko, baina baliteke emaitza onak lortzea, eta oso interpretagarriak.

Azkenik, *emojien* erabilera aztertzea interesgarria litzateke, mezuen analisi linguistiko zein sailkapenari begira. Izan ere, elementu hauek sare sozialetan funtzio linguistikoak, sintaktiko, semantiko zein pragmatikoak dituztela ikusi da, idatzitakoari konnotazio jakin bat emateko edota sentimendu zein emozioak era bisualean adierazteko erabiltzen baitira [64]. Beraz, mezuen ezaugarri linguistiko gisa azter daitezke, baita sailkapenean iragarpenak egiteko ezaugarri iragarle gisa erabili ere.

7.4 Ondorio pertsonalak

Proiektu hau azken urtean izan dudana garapenaren islada da. Gai honekin lehenengo kontaktua 2022ko udan izan nuen, IXA taldean praktikak egitean. Une hartan ez nekien ezer hizkuntzaren prozesamenduaren arloari buruz. Apurka-apurka, esparru honetan nuen jakintza zabaltzen joan naiz, bai praktiketan ikasitakoari, bai Hizkuntzaren Prozesamendua eta Ikasketa Automatikoa eta Neurona-Sareak irakasgaiei, bai Gradu Amaierako Lan honi esker.

Graduan zehar eskuratutako ezagutzaz gain, proiektu honen garapenean hainbat kontzeptu, prozesu, teknika eta tresna berri erabiltzen ikasi dut. Batetik, aurrekarien analisia egiten ikasi dut, artikulua ezberdinak irakurri eta informazioa sintetizatzen, hauen inguruko azalpenak ematen testu zein taulen bidez eta erreferentziak era zuzenean egiten. Bestetik, GPUak eta hauek atzitzeko erabiltzen diren ilara-sistemak erabiltzen trebatu naiz, IXA taldeko makinekin. Kontzeptu teorikoei dagokienean, topiko-ereduen berri izan dut eta hauekin erlazioatutako kontzeptuak barneratu eta praktikara eramateko aukera izan dut.

Bestetik, nahasmendu psikologikoei buruz nuen ezagutza zabaltzeko aukera izan dut irakurritako hainbat artikuluri esker, baita hauen ikerkuntzan sare sozialek duten garrantziaren berri izateko aukera ere. Aurrekariak aztertzeari esker hainbat nahasmenduren detekzioan egindako saiakerak ezagutu ditut, baita aurrerakuntzak bultzatzeko honen inguruan antolatzen diren nazioarteko erronkak ere.

Gainera, orain arte egindako lanen aldean luzeagoa den proiektu bat kudeatzen ikasi dut, partaideen arteko komunikazioaren garrantziaz ohartu naiz eta memoria bat era argi eta antolatuan idazteak suposatzen duen lanaz jabetu naiz.

Honetaz gain, IkerGazte euskal ikertzaile gazteen kongresuan ekarpen bat egiteko aukera izan dut, artikulua bat argitaratuz eta lana kongresuan bertan aurkeztuz. Esperientzia guztiz berria izan da niretzat, oso aberasgarria.

Hurbilpen oinarritzkoagoaren esperimentazioa

Eranskin honetan sailkapena egiteko erabilitako hurbilpen oinarritzkoagoarekin lortutako emaitzak aurki daitezke. Hurbilpen honetan topiko kopuru (m) ezberdinekin osatutako errepresentazioak erabili dira eta lan honetan proposatutako oinarri-lerro zein ikasketa automatikoko hainbat ereduren bitartez sailkatu dira (informazio gehiago [6.4. atalean](#)). Kasu guztietan lortutako emaitzak [A.1.](#) taulan bildu dira.

m	Sailkatzailea	Asmatze-tasa	Doitasuna	Estaldura	F puntuazioa
2	Oinarri-lerroa	82,9	70,7	95,4	81,2
	Random Forest	81,4	75,9	76,3	76,1
	Bagging	87,5	82,2	86,7	84,4
	XGBoost	87,4	82,2	86,4	84,2
5	Oinarri-lerroa	81,9	72,5	85,8	78,6
	Random Forest	86,9	82,2	84,8	83,5
	Bagging	87,0	81,8	85,5	83,6
	XGBoost	87,2	81,9	86,2	84,0
10	Oinarri-lerroa	82,1	72,7	86,5	79,0
	Random Forest	86,2	79,8	86,6	83,1
	Bagging	87,0	82,5	84,4	83,5
	XGBoost	85,8	79,2	86,2	82,5
50	Oinarri-lerroa	81,4	78,8	71,4	74,9
	Random Forest	81,1	70,5	88,6	78,5
	Bagging	87,1	85,7	80,3	82,9
	XGBoost	85,4	79,9	83,3	81,6
100	Oinarri-lerroa	76,0	70,4	66,2	68,2
	Random Forest	77,6	65,5	89,7	75,8
	Bagging	86,7	84,6	80,5	82,5
	XGBoost	84,9	78,4	84,6	81,4
150	Oinarri-lerroa	74,3	67,1	66,6	66,9
	Random Forest	77,2	64,8	90,4	75,5

A. HURBILPEN OINARRIZKOAGOAREN ESPERIMENTAZIOA

m	Sailkatzailea	Asmatze-tasa	Doitasuna	Estaldura	F puntuazioa
	Bagging	87,3	85,3	81,3	83,3
	XGBoost	85,7	81,8	81,2	81,5
200	Oinarri-lerroa	73,0	65,4	64,8	65,1
	Random Forest	80,9	70,0	89,2	78,4
	Bagging	88,1	86,8	82,0	84,3
	XGBoost	87,2	84,1	82,9	83,5
250	Oinarri-lerroa	76,0	71,3	64,0	67,5
	Random Forest	81,1	70,1	89,5	78,6
	Bagging	88,7	87,0	83,3	85,1
	XGBoost	87,3	82,9	84,8	83,8
300	Oinarri-lerroa	74,2	68,5	62,4	65,3
	Random Forest	80,8	69,4	90,5	78,6
	Bagging	88,9	87,4	83,4	85,4
	XGBoost	87,7	83,8	84,7	84,2

A.1 Taula: Topiko kopuru (m) ezberdineko LDA errepresentazioen gainean sailkatzaile klasikoekin lortutako emaitzak (%).

Bibliografia

- [1] Anna Kolliakou, Ioannis Bakolis, David Chandran, Leon Derczynski, Nomi Werbeloff, David PJ Osborn, Kalina Bontcheva, and Robert Stewart. Mental health-related conversations on social media and crisis episodes: A time-series regression analysis. *Scientific reports*, 10(1):1–7, 2020. Ikusi 1 orrialdea.
- [2] Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological bulletin*, 143(2):187, 2017. Ikusi 1 orrialdea.
- [3] Matthew K Nock, Franchesca Ramirez, and Osiris Rankin. Advancing our understanding of the who, when, and why of suicide risk. *JAMA psychiatry*, 76(1):11–12, 2019. Ikusi 1 orrialdea.
- [4] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860, 2018. Ikusi 2 orrialdea.
- [5] Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33, 2019. Ikusi 2, 12, and 17 orrialdeak.
- [6] Roger E Meyer, Carl Salzman, Eric A Youngstrom, Paula J Clayton, Frederick K Goodwin, J John Mann, Larry D Alphs, Karl Broich, Wayne K Goodman, and John F Greden. Suicidality and risk of suicide—definition, drug safety concerns, and a necessary target for drug development: a consensus statement. *The Journal of clinical psychiatry*, 71(8):20326, 2010. Ikusi 2 orrialdea.
- [7] Alex Crosby, LaVonne Ortega, and Cindi Melanson. Self-directed violence surveillance; uniform definitions and recommended data elements. Technical report, US Centers for Disease Control and Prevention, 2011. Ikusi 2 orrialdea.
- [8] Rebecca A Bernert, Melanie A Hom, and Laura Weiss Roberts. A review of multidisciplinary clinical practice guidelines in suicide prevention: toward an emerging standard in suicide risk assessment and management, training and practice. *Academic Psychiatry*, 38:585–592, 2014. Ikusi 2 orrialdea.
- [9] Joseph H Obegi. Rethinking suicidal behavior disorder. *Crisis*, 2018. Ikusi 2 orrialdea.
- [10] Bonnie Harmer, Sarah Lee, TVH Duong, and Abdolreza Saadabadi. *Suicidal ideation*. StatPearls Publishing, 2023. Ikusi 2 orrialdea.
- [11] Kathryn Downey Piscopo. Suicidality and death by suicide among middle-aged adults in the United States. *The CBHSQ Report*, 2017. Ikusi 2 orrialdea.
- [12] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39, 2015. Ikusi 11, 12 orrialdeak.

- [13] Adrienne S Juarascio, Amber Shoab, and C Alix Timko. Pro-eating disorder communities on social networking sites: a content analysis. *Eating disorders*, 18(5):393–407, 2010. Ikusi 11 orrialdea.
- [14] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 579–582, 2014. Ikusi 11 orrialdea.
- [15] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1171–1184, 2016. Ikusi 11 orrialdea.
- [16] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International conference on web search and data mining*, pages 91–100, 2017. Ikusi 11 orrialdea.
- [17] Stevie Chancellor, Zhiyuan Lin, and Munmun De Choudhury. “This post will just get taken down” Characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1157–1162, 2016. Ikusi 11 orrialdea.
- [18] Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H Andrew Schwartz. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, 2018. Ikusi 12 orrialdea.
- [19] Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, 2022. Ikusi 12 orrialdea.
- [20] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110, 2016. Ikusi 12 orrialdea.
- [21] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697, 2020. Ikusi 12 orrialdea.
- [22] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010. Ikusi 13, 23, 33, 37, and 58 orrialdeak.
- [23] Kyungil Kim, Sunhee Choi, Joohyung Lee, and Jonghan Sea. Differences in linguistic and psychological characteristics between suicide notes and diaries. *The Journal of general psychology*, 146(4):391–416, 2019. Ikusi 13, 33, 34, and 37 orrialdeak.
- [24] Bridianne O’dea, Mark E Larsen, Philip J Batterham, Alison L Calear, and Helen Christensen. A linguistic analysis of suicide-related Twitter posts. *Crisis*, 2017. Ikusi 13, 31, 33, 34, and 37 orrialdeak.
- [25] M Johnson Vioules, Bilel Moulahi, Jérôme Azé, and Sandra Bringay. Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, 62(1):7–1, 2018. Ikusi 13, 16, 33, and 37 orrialdeak.
- [26] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7, 2019. Ikusi 13, 14, 32, 33, and 37 orrialdeak.

-
- [27] Tatiana A Litvinova, Pavel V Seredin, Olga A Litvinova, and Olga V Romanchenko. Identification of suicidal tendencies of individuals based on the quantitative analysis of their internet texts. *Computación y Sistemas*, 21(2):243–252, 2017. Ikusi 13 orrialdea.
- [28] Bridianne O’Dea, Tjeerd W Boonstra, Mark E Larsen, Thin Nguyen, Svetha Venkatesh, and Helen Christensen. The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study. *Plos one*, 16(5):e0251787, 2021. Ikusi 13 orrialdea.
- [29] Asma Abdulsalam and Areej Alhothali. Suicidal ideation detection on social media: A review of machine learning methods. *arXiv preprint arXiv:2201.10515*, 2022. Ikusi 15, 16, and 24 orrialdeak.
- [30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. Ikusi 15, 24 orrialdeak.
- [31] Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018, 2018. Ikusi 15 orrialdea.
- [32] Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: student research workshop*, pages 147–156, 2019. Ikusi 15 orrialdea.
- [33] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. Tracking suicide risk factors through Twitter in the US. *Crisis*, 2014. Ikusi 15, 16 orrialdeak.
- [34] Peter M Lewinsohn, Paul Rohde, and John R Seeley. Psychosocial risk factors for future adolescent suicide attempts. *Journal of consulting and clinical psychology*, 62(2):297, 1994. Ikusi 15 orrialdea.
- [35] Reilly N. Grant, David Kucher, Ana M. León, Jonathan F. Gemmell, Daniela S. Raicu, and Samah J. Fodeh. Automatic extraction of informal topics from online suicidal ideation. *BMC bioinformatics*, 19(8):57–66, 2018. Ikusi 15, 16 orrialdeak.
- [36] Samah Fodeh, Taihua Li, Kevin Menczynski, Tedd Burgette, Andrew Harris, Georgeta Ilita, Satyan Rao, Jonathan Gemmell, and Daniela Raicu. Using machine learning algorithms to detect suicide risk factors on Twitter. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 941–948. IEEE, 2019. Ikusi 15, 16 orrialdeak.
- [37] John A Hartigan and Manchek A Wong. Algorithm AS 136: A k-Means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. Ikusi 15 orrialdea.
- [38] Syed Tanzeel Rabani, Qamar Rayees Khan, and AMUD Khanday. Detection of suicidal ideation on Twitter using machine learning & ensemble approaches. *Baghdad Science Journal*, 17(4):1328–1339, 2020. Ikusi 15 orrialdea.
- [39] Gayathri Ananthkrishnan, Ashok Kumar Jayaraman, Tina Esther Trueman, Satanik Mitra, AK Abinesh, and Abirami Murugappan. Suicidal intention detection in tweets using BERT-based transformers. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 322–327. IEEE, 2022. Ikusi 16 orrialdea.
- [40] *Reddit C-SSRS Suicide Dataset*. Zenodo, May 2019. Ikusi 17 orrialdea.
- [41] Milan Straka, Jan Hajic, and Jana Straková. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4290–4297, 2016. Ikusi 22, 58 orrialdeak.

- [42] Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October 2018. Association for Computational Linguistics. Ikusi 22 orrialdea.
- [43] Vatsal. Word2Vec explained. Towards Data Science, 2021. Ikusi 24 orrialdea.
- [44] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. Ikusi 25 orrialdea.
- [45] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012. Ikusi 25 orrialdea.
- [46] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020. Ikusi 25 orrialdea.
- [47] M Blei David and D Lafferty John. Topic models. *Text Mining*, pages 101–124, 2009. Ikusi 25 orrialdea.
- [48] Yuanhan Mo, Georgios Kononatsios, and Sophia Ananiadou. Supporting systematic reviews using LDA-based document representations. *Systematic reviews*, 4(1):1–12, 2015. Ikusi 25 orrialdea.
- [49] Jorge Pérez, Alicia Pérez, Arantza Casillas, and Koldo Gojenola. Cardiology record multi-label classification using Latent Dirichlet Allocation. *Computer methods and programs in biomedicine*, 164:111–119, 2018. Ikusi 25 orrialdea.
- [50] Matthew Hoffman, Francis Bach, and David Blei. Online learning for Latent Dirichlet Allocation. *advances in neural information processing systems*, 23, 2010. Ikusi 26 orrialdea.
- [51] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013. Ikusi 26 orrialdea.
- [52] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001. Ikusi 27 orrialdea.
- [53] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996. Ikusi 27 orrialdea.
- [54] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. Ikusi 27 orrialdea.
- [55] Nitin Kumar Chauhan and Krishna Singh. A review on conventional machine learning vs deep learning. In *2018 International conference on computing, power and communication technologies (GUCON)*, pages 347–352. IEEE, 2018. Ikusi 27 orrialdea.
- [56] Dan Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009. Ikusi 27 orrialdea.
- [57] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. Ikusi 28 orrialdea.
- [58] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. Ikusi 28 orrialdea.
- [59] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. Ikusi 28 orrialdea.
- [60] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. Ikusi 28 orrialdea.

- [61] Anoop Sarkar. The conflict between future tense and modality: The case of will in English. *University of Pennsylvania Working Papers in Linguistics*, 5(2):6, 1998. Ikusi [35](#) orrialdea.
- [62] Shraddha Anala. A guide to word embeddings. Towards Data Science, 2020. Ikusi [44](#) orrialdea.
- [63] Sarojadevi Palani, Prabhu Rajagopal, and Sidharth Pancholi. T-BERT–Model for sentiment analysis of micro-blogs integrating topic model and BERT. *arXiv preprint arXiv:2106.01097*, 2021. Ikusi [60](#) orrialdea.
- [64] Muhammad Hasyim. Linguistic functions of emoji in social media communication. *Opcion*, 35, 2019. Ikusi [60](#) orrialdea.