V. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2023ko maiatzaren 17, 18 eta 19a
Donostia, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Ideiagintza suizidaren identifikazioa sare sozialetan

Sara Gracia,
Maite Oronoz Antxordoki
eta *Alicia Pérez Ramírez*

123-130 or.
https://dx.doi.org/10.26876/ikergazte.v.03.16

Ideiagintza suizidaren identifikazioa sare sozialetan

Sara Gracia[1], Maite Oronoz[2], Alicia Pérez[2]

[1] Informatika Fakultatea (UPV/EHU). M. Lardizabal 1. 20080 Donostia.
[2] Ixa ikerketa taldea (UPV/EHU). HiTZ: Basque Center for Language Technology. M. Lardizabal 1. 20080 Donostia.
maite.oronoz@ehu.eus alicia.perez@ehu.eus

sgracia003@ikasle.ehu.eus

Laburpena

Suizidioa gizartearen kezka nagusietako bat bilakatu da azken urteetan. Gainera, sare sozialak gure egunerokotasunaren parte bilakatu dira, eta emozioak adierazteko erabiltzen dira askotan. Lan honetan sailkapen bitarra burutu da Reddit sare sozialeko mezu baten edukia suizidioarekin erlazionatua dagoen ala ez erabakitzeko. Alde batetik, artearen egoerako sistema gainbegiratuei dagokienean, ELECTRA transformerrarekin lortu da asmatze-tasarik altuena, %97,9koa. Bestalde, ondorioztatu da LDA topiko-ereduak sortutako errepresentazioak baliagarriak izan daitezkeela ataza honetan, eta hau frogatzeko oinarri-lerroa den sailkatzailea proposatu da, %83,3ko asmatze-portzentaia izan duena 5 topiko erabiliz.

Hitz gakoak: hizkuntzaren prozesamendua, suizidio ideiagintzaren detekzioa, topiko-ereduak, ikasketa sakona

Abstract

Suicide has become one of society's main concerns in recent years. In addition, social media has become part of our everyday life and is often used to express emotions. In this work, a binary classification has been carried out to determine whether or not the content of a message on the Reddit social network is related to suicide. On the one hand, with regard to supervised systems in the state of the art, the best performance has been achieved with the ELECTRA transformer, with an accuracy-rate of 97.9%. On the other hand, it has been concluded that the representations produced by the LDA topic-model can be useful for this task, and to prove this, a baseline classifier has been proposed, which has reached an accuracy of 83.3%.

Keywords: *natural language processing, suicidal ideation detection, topic models, deep learning*

1 Sarrera eta motibazioa

Sare sozialetan hainbat nahasmendu psikologikorekin lotutako komunitateak daude. Horietan, milaka pertsonek beren egoera emozionala partekatzen dute, laguntza eskatzen dute edo, besterik gabe, txateatu egiten dute. Nature aldizkarian argitaratutako ikerketa-lan batek ondorioztatu zuen buruko nahasmenduak dituzten pazienteek osasunzerbitzuei egindako premiazko laguntza-eskaerak ugaritu egin zirela osasun mentalarekin lotutako gaiei buruzko txio gehien argitaratu ziren egunetan (Kolliakou et al., 2020).

Askotan, suizidioa buru-nahasmendu batek eragindakoa izaten da. Osasunaren Mundu Erakundearen (OME) arabera, suizidioa gazteen arteko (15-24 urte) heriotza-kausa ohikoenetako bat da (Vargas eta Saavedra, 2012). Azken aldian, badira hizkuntzaren prozesamenduaren inguruko teknikak erabiliz ideiagintza suizidaren identifikazioa ardatz duten ikerketak. Lan honetan, Reddit (www.reddit.com) sare sozialetik eskuratutako mezu sorta batean sailkapen bitarreko ikuspuntutik zantzu suizidak erakusten dituzten mezuak identifikatzeko metodo ezberdinak landu ditugu. Lan honen ekarpen nagusia honakoa da: artearen egoeran dauden eta makina-baliabide handiak behar dituzten sistema batzuk berrerabili eta erreproduzitu ditugu eta hauen emaitzak oinarrizkoagoa den hurbilpen batekin konparatu ditugu (LDA topiko-ereduak sortutako errepresentazioa baliatzen duen sistema batekin).

Jarraian datozen ataletan honako edukia jorratu dugu: 2. atalean aurrekarietan burututako lanak aipatzen dira. Ondoren, 3. atalean, ikerketaren muina deskribatzen da, erabilitako datuak eta sistemak azalduz eta lortutako emaitza esperimentalak erakutsiz. 4. atalean esperimentuetatik ateratako ondorioak aurki daitezke eta, amaitzeko,

5. atalean lan hau hedatzeko etorkizunerako norabideen proposamenak burutzen dira.

2 Arloko egoera eta ikerketaren helburuak

Sare sozialak abiapuntutzat hartuz suizidio ideiagintza detektatzeko egindako saiakerak nabarmenki ugaritu dira azken aldian, arazo honek gaur egungo gizartean duen garrantziagatik. Ildo horretan, CLPsych (Zirikly et al., 2019) bezalako nazioarteko erronkak burutu dira ikerketa arlo honetan aurrerapenak sustatzearren. Abdulsalam eta Alhothali autoreek (2022) ikerketan gai hau jorratzen duten lan asko bildu eta azaltzen dituzte. Egile horiek egindako berrikuspen bibliografikoan aztertutako lan gehienetan, Word2Vec (Mikolov et al., 2013) gisako hitz-embeddingak, n-gramak, edota hitzen agerpen-maiztasunekin zerikusia duten metrikak, TF-IDF eta LIWC esaterako, erabiltzen dira sare sozialetako mezuak era numerikoan errepresentatzeko. Helburu honekin topiko-ereduak erabiltzen dituzten lanak, aldiz, gutxi dira eta hori izan da, besteak beste, aztertu nahi izan dugun errepresentaziorako modu bat.

Topiko-eredua ikasketa automatikoko teknika da, dokumentu bilduma batean egitura semantiko ezkutuak identifikatzeko gai dena (Blei et al., 2003; Blei, 2012; Dieng et al., 2020). Hauek, testuen dimentsionalitate baxuko errepresentazioak lortzeko balio dezakete. Fodeh et al. autoreek (2019) lanean LDA (*Latent Dirichlet Allocation*) topiko-eredua erabiltzen dute txioak errepresentatzeko. Gainera, autoreak gai izan dira Jashinsky et al. autoreek (2014) lanean proposatutako 12 suizidio-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 arren, autoreek ondorioztatzen dute K-Means multzokatze-algoritmo ez-gainbegiratua ez dela egokia ataza hori ebazteko.

Erabilitako sailkatzaileei dagokienean, gehien erabiltzen diren sailkatzaile klasikoak SVM, Random Forest eta erregresio logistikoa dira. Ikasketa sakonean oinarritutakoen artean berriz, LSTM eta CNN sare-neuronalak dira erabilienak. Azken urteetako lan batzuek transformerrak erabiltzen dituzte, Ananthakrishnan et al. autoreen (2022) lana kasu. Bertan, Google enpresak sortutako *Bidirectional Encoder Representations from Transformers* (BERT) sailkatzailearen aldaera ezberdinak erabiltzen dira txio batean ideiagintza suizidarik dagoen ala ez erabakitzeko, %95,4ko asmatze-tasa lortuz. BERT transformerretan oinarritutako hizkuntza-ereduek lortzen dituzten asmatze-tasa oso altu horiek direla-eta, hurbilpen hau bereziki aztertu nahi izan dugu 3.2.1 atalean.

Lan honen helburua da sare sozialetako mezuak bi klasetan sailkatzeko, suizidio zantzuak dituztenak eta ez dituztenak, hurbilpen *'garestiak'* eta hurbilpen *'oinarrizkoagoak'* konparatzea. Horretarako, datu-sorta beretik abiatuta bi saiakera ezberdin egin ditugu eta ondoren, hauekin lortutako emaitzak alderatu: Alde batetik, artearen egoerako ikasketa sakoneko sailkatzaileak erabili ditugu mezuak sailkatzeko, mezuen errepresentazio gisa hitz-embeddingak erabiliz. Bestalde, errepresentazio numeriko berri bat sortu dugu topiko-ereduak erabiliz, eta oinarri-lerroa litzatekeen sailkatzailea definitu dugu.

3 Ikerketaren muina

Aurrekariak aztertuta, gure **hipotesia** ideagintza suizida era ez-gainbegiratuan inferitutako topikoen baitan islatu litekeela da; eta are gehiago, topikoetatik bertatik iragar daitezkeela ideagintza suizidaren zantzuak. Gure lanaren **sendotasuna**, datu-sorta beretik abiatuta sailkapen bitarra burutzeko metodo ezberdinen azterketan datza. Gainera, topiko-ereduen gisako metodo ez-gainbegiratuek eduki semantikoa era numerikoan biltzeko izan dezaketen gaitasuna ebaluatu dugu. Hizkuntzaren ulermen artifiziala da lan honen esparrua, izan ere, testu-sortetatik informazioa erauzteko gaitasuna dituzten algoritmoez baliatuko gara.

Jarraian, 3.1. atalean, atazarako eskuratu dugun testu-sorta eta honen prestaketa azaldu dugu. Ondoren, 3.2. atalean, lan honetan proposatutako metodologia bildu dugu, inplementatu eta probatu ditugun hurbilpen ezberdinak hain zuzen ere. 3.2.1. atalean ikerketa honetan erabilitako artearen egoerako sailkatzaile gainbegiratuak deskribatzen dira, eta hauekin lortutako emaitzak azaldu. 3.2.2. atalean berriz, topiko-ereduak erabiliz sortutako errepresentazio numerikoa proposatzen da, eta honek sailkapen ataza batean izango lukeen oinarri-lerroa definitu.

3.1 Erabilitako datuak eta burututako prestaketa

Lan honetan, *Suicide and Depression Detection*[1] deritzon corpusa erabili da. Hau, Reddit sare sozialeko "SuicideWatch" eta "depression" *subreddit* edo foroetatik eskuratutako mezu bilduma da. "SuicideWatch" foroko mezuak

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

2008/12/16tik 2021/01/02ra bitartekoak dira eta "depression" forokoak, aldiz, 2009/01/01-2021/01/02 tartekoak. Hizkuntzari dagokionean, mezuak ingelesez idatziak izan dira. Guztira, 232.704 mezu biltzen dira. Mezuek etiketa bana dute esleituta, alegia, klasea. Klaseak har ditzakeen balioak bi dira: *suicide*, mezuak suizidio zantzuak dituela adierazteko; *non-suicide*, suizidioarekin erlazionatuta ez dauden mezuei lotuta. Klase banaketari dagokionean, banaketa uniformea da.

Sare sozialetako testua, normalean, desegituratuagoa izaten da beste iturri batzuetatik eskuratutakoa baino, testuinguru informaltzat hartzen baita. Arrazoi honengatik, datuen prestaketa oso garrantzitsua izaten da. Hau burutzeko erabilitako prozesua Suicidal-BERT[2] proiektuko autoreek erabilitakoaren erreprodukzioa da. Bertan, prestaketa bi etapetan banatzen da: aurreprozesaketa eta garbiketa.

Aurreprozesaketan hainbat urrats ezberdin eman dira. Hauen helburu nagusia hiztegiaren tamaina txikitzea izan da, datuen dimentsionalitatea murriztu eta ereduen 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:

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 guzti hauek garrantzia berezirik ez dutenez, ezabatzea erabaki da.
5. Hitzen luzera mugatu: Datu-sorta honetako hainbat mezuetan era okerrean errepikatutako hizkiak aurkitu dira, *goooood* 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 tankerako testuinguru informaletan, ez zaio hainbesteko garrantzia ematen zuzentasun ortografikoari. Aurki daitezken akatsak zuzentzen saiatzeko asmoz, Pythoneko *Symspell* algoritmoa erabili da.
7. Hitz-hutsak (*Stopwords*) ezabatu: Hitz-hutsak testuetan askotan erabiltzen diren baina esanahi berezirik ez duten hitzak dira, "eta" bezalako lokailuak, esaterako. Informaziorik ematen ez dutenez, hauek ezabatzea erabaki da. Ezeztapena adierazten duten ingeleseko *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. Lematizazioa: Lematizazioa analisi morfologikoaren bidez hitzen lemak eta morfemak lortzeko prozesua da. Hitz guztiak beren lemengatik ordezkatu dira.

Aurreprozesaketaren ostean **garbiketa** egin da. Kasu honetan, pauso hauek jarraitu dira:

1. *"filler"* hitzaren agerpenak ezabatu: Hitz honek agerpen kopuru handia izan arren, testuinguru honetan esanahi konkreturik ez duenez, ezabatu egin da.
2. Testu hutsak ezabatu: Aurreprozesaketaren ostean testu asko hutsak bilakatu direnez, ezabatu egin dira.
3. Mezu luzeak ezabatu: 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.

Jatorrizko datu bildumaren, aurreprozesaketaren, eta garbiketaren ostean lortutako mezuen luzerarekin lotutako estatistikoak 1a taulan ikus daitezke. Mezuen klase-banaketa aldiz, 1b taulan dago bilduta.

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

[2]Proiektuaren kodea eskuragarri hemen: `https://github.com/gohjiayi/suicidal-text-detection/`

batezbeste 132 hitz zituen eta aurreprozesaketaren eta garbiketaren ondoren, 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 uniformea hautsi da, *non-suicide* klaseko mezu portzentaia %61,1ekoa izanik, *suicide* klasekoena %38,9koa den bitartean.

1. Taula: **Datu-sortaren deskribapena: (a) Mezuen hitz kopurua (b) Mezuen klaseak. Datuak hiru zutabetan banatzen dira, bakoitza prestaketako fase bati dagokiolarik: Jatorrizko corpusa; aurreprozesaketa eta gero; garbiketa eta gero.**

(a) **Mezuen hitz kopuruarekin lotutako estatistikoak, non $\bar{x}$ batezbestekoa den, σ desbideratze-estandarra eta %25, %50 eta %75 berriz, kuartilak.**

	Corpusa	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

(b) **Mezuen klase-banaketa**

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

Aurreprozesatutako eta garbitutako *Suicide and Depression Detection* corpusaren klase-banaketa errespetatuz, hau hiru azpimultzotan banatu da: entrenamenduko azpimultzoa, eredua entrenatzeko erabiltzen dena (%80); garapenekoa, entrenamendu prozesuan eredua ebaluatzeko erabiltzen dena (%10); eta probakoa, entrenamenduaren ostean eredua ebaluatzeko erabiltzen dena (%10).

Aurreprozesatutako eta garbitutako corpusaren aberastasunaren berri izatearren, entrenamendu multzoko hiztegiaren tamaina nabarmendu nahi dugu, 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 (OOV, *out-of-vocabulary*) kopurua da. Zehazki, garapen eta proba azpimultzoetan agertutako hitzak zeinak entrenamendu multzoan agertzen ez diren. Garapen azpimultzoaren kasuan hauen kopurua 1.122 da, azpimultzo honetako hitzen %8,55; eta probaren kasuan, berriz, 1.163, azpimultzoko hiztegiaren %8,89. Entrenamendu multzoan mezu asko egon arren, testu berriak prozesatzean hitz berriak agertzeko parada handia dela ikusten dugu. Laburtuz, corpusa lexikoki aberatsa da eta entrenamenduko hiztegitik at dauden hitzak azpimultzo bakoitzeko %8 dira.

3.2 Sailkapen bitarra burutzeko erabilitako metodoak

3.2.1 Ikasketa sakonean oinarritutako sailkatzaile gainbegiratuak

Ikasketa gainbegiratuan datu etiketatuak erabiltzen dira. Era honetako ereduek datu eta etiketen arteko erlazioak ikasten dituzte helburu ezberdinekin. Horietako bat sailkapena da, lan honetan ebatzi nahi den ataza. Kasu honetan, datuak Reddit sare sozialeko mezuak dira eta etiketak edo klaseak berriz, mezu bakoitza suizidioarekin erlazionatuta dagoen ala ez adierazten duena. Helburua ereduak etiketatutako adibideetatik ikastea da, ondoren, mezu bat jasota suizidio zantzuak dituen ala ez erabakitzeko gai izateko. Kasu honetan, ataza bera ebazten duen Suicidal-BERT proiektuan eraikitako bost eredu gainbegiratu erreproduzitu dira, ikerketa arlo honen artearen egoera definitzen dutenak:

- Erregresio logistikoa: Metodo estatistiko hau hizkuntzaren prozesamenduan sailkapenerako oinarri-lerro gisa erabili ohi da (Jurafsky eta Martin, 2009).
- Sare neuronal konboluzionala (CNN), (LeCun et al., 1995): Sare neuronal mota hauek konboluzioetan oinarritzen dira, datuetatik ezaugarri edo patroiak erauzteko gai direnak. Ikusmen artifizialean erabiltzen dira

gehien bat, baina badaude testuekin lan egiteko prestatutako ereduak ere.

- LSTM eredua (Hochreiter eta Schmidhuber, 1997): Neurona-sare errekurrente (RNN) mota bat da, desagertzen diren gradienteen arazoa konpontzen duena. Testu, audio edota bideo moduko datu sekuentzialetan erabiltzen dira gehien bat.
- BERT transformerra (Devlin et al., 2019): Googlek sortutako eredua da, atentzio-mekanismoa erabiltzen duena. Gaur egun oso erabilia da eta tamaina ezberdinetako bertsioak daude eskuragarri. Kasu honetan, BERT-base deritzona erabili da.
- ELECTRA transformerra (Clark et al., 2020): Hau ere Googlek sortutako transformer eredua da. BERTekin alderatuta, entrenamendu datu gutxiago behar ditu errendimendu bera lortzeko, adibide bakoitzeko informazio gehiago jasotzeko gaitasuna baitu. Oraingoan ere, ELECTRA-base eredua erabili da bi transformerren arteko alderaketa egokiena egiteko asmoz.

Aipatutako lehen hiru ereduen sarrera gisa prestatutako mezuen hitz-embeddingak erabili dira. Hitz-embeddingak hitzen zenbakizko errepresentazio sakonak dira, hau da, hitzak zenbakizko bektore bihurtzen dituzte. Kasu honetan, datu-sortaren gainean entrenatutako Word2Vec eredua erabiliz lortu dira bektore horiek (Mikolov et al., 2013). Ondoren, hitz-embedding horiek erregresio logistikoaren eta CNN eta LSTM ereduen sarrera gisa erabili dira. BERT eta ELECTRA transformerren kasuetan aldiz, mezu originalak erabili dira sarrera bezala, aurreprozesatu gabeak. Izan ere, eredu hauek beren aurreprozesaketa burutzen dute eta baita beren hitz-embedding propioak kalkulatu ere.

Aurkeztutako 5 eredu gainbegiratuekin lortutako emaitzak 2 taulan ikus daitezke. Erabilitako metrika guztietan gailendu den eredua ELECTRA izan da, %97,9ko asmatze-tasarekin. Honen emaitzetan sakontzeko asmoz erroreen azterketa burutu da. NF (negatibo-faltsu) erroreek suizidio zantzuak dituzten eta sistemak hala identifikatu ez dituen mezuak errepresentatzen dituzte. PF (positibo-faltsu) erroreak aldiz aurkako kasuetan gertatzen dira, benetan suizidioarekin zerikusirik ez dituzten mezuak sistemak gai horrekin erlazioa duten mezu gisa sailkatzen dituenean. Ataza honetan NF errorea minimizatzea komeni da. ELECTRA transformerrarekin lortutako emaitzen kasuan, NF errore kopurua 158 mezukoa izan da, PF erroreena handiagoa izan den bitartean, 204 mezukoa hain zuzen ere. Jarraian, errore bakoitzaren adibide bana aurkezten da:

- 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.*.

2. Taula: **Eredu gainbegiratuekin lortutako emaitzak (%)**

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**

3.2.2 Topiko-ereduak erabiliz lortutako errepresentazioa eta proposatutako oinarri-lerroa

Topiko-eredua ikasketa automatikoko teknika da, dokumentu bilduma batean egitura semantiko ezkutuak identifikatzeko gai dena (Blei et al., 2003; Blei, 2012; Dieng et al., 2020). Hauek inplementatzeko teknika erabilienetako bat *Latent Dirichlet Allocation* (LDA) da, probabilitate-eredu Bayestar hierarkikoa (Blei et al., 2003). Topiko bakoitza hitzen gaineko probabilitate-banaketa gisa errepresentatzen du, eta dokumentu bakoitza topikoen banaketa gisa. Dokumentu multzo baten gainean aplikatzean, dokumentu bakoitzaren topikoen probabilitate-banaketak testuaren dimentsionalitate baxuko errepresentazioa eskaintzen du (Hoffman et al., 2010, 2013). Lan honetan, mezuen **errepresentazio** numeriko bat lortzeko baliatu da LDA, hain zuzen ere.

Har dezagun ondorengo mezua adibide gisa: *"It ends tonight.I can't do it anymore. I quit."*. Mezu honen LDA errepresentazioa, topiko kopurua $m = 5$ izanik, honakoa da: $\mathbf{x}_i = (0,033; 0,034; 0,033; 0,033; 0,867)$. Bost luzerako bektore horretan, balio bakoitzak mezu horrek topiko batekiko duen erlazioaren proportzioa adierazten

du. Adibideko kasuan 5. topikoa gailentzen da, mezuak erlaziorik handiena topiko horrekin duela adieraziz. Balio hauen interpretazioan laguntzeko, topiko bakoitzari dagokion hitz-hodeia aurki daiteke 1 irudian. Bertan, topikoei dagozkien hitz-banaketak adierazten dira, non hitzaren tamaina honek topikoan duen garrantziarekiko proportzionala den. Esan bezala, landutako adibidean 5. topikoa da nagusi. Honen hitz-hodeiari erreparatuta, mezuan agertzen den "*anymore*" aurki dezakegu, baita "*not*" ere, *can't* kontrakzioaren parte dena. Gogoratu, testuari aurreprozesatzea aplikatzean, kontrakzioak hedatu direla. Topikoak aurreprozesatutako eta garbitutako corpusarekin kalkulatzen dira.

Hitz-hodeiei dagokienean, 1. topikoan ikastetxeekin eta honek Redditeko erabiltzaileengan sortzen dituen erreakzio eta emozioekin erlazionatutako hitzak aurki daitezke: "eskola", "joan", "kaka", "izorratu", "klasea", ... 2. topikoan aldiz, "jakin", "ez", "atsegin", "laguna", "hitz egin" eta antzeko hitzak agertzen dira. 3. topikoan berriz, "atsegin', "nahi', "ikusi', "jolastu' eta "jokoa' bezalako hitzak. 4. topikoan Reddit sare sozialarekin erlazionatutako lexikoa aurki daiteke, "*geddit*", "*post*" eta "*sub*", esaterako. 5. topikoan gailentzen diren hitzak "ez", "bizi", "eraildu", "sentitu" eta "bizitza" gisakoak dira.

1. Irudia: $m = 5$ **kasuan lortutako topikoen hitz-hodeiak**

1. topikoa

2. topikoa

3. topikoa

4. topikoa

5. topikoa

Azpimarratu beharrekoa da LDA algoritmoaren aldaera ez-gainbegiratua erabili dela, alegia, ereduak ez duela corpusean jorratzen diren gaiei buruzko aldez aurreko zantzurik. Izan ere, topiko edo gaiak automatikoki aurkitzen dira bildumako testuak aztertuta. Arrazoi hauengatik, topiko-ereduak datu-bilduma ez-egituratuetan nolabaiteko antolamendua aurkitzeko metodo baliagarri bilakatu dira (Blei David eta John, 2009; Mo et al., 2015).

Teknika ez-gainbegiratu hau erabilita probak egin dira topiko kopuru (m) ezberdinekin, honek jatorrizko mezuen adierazpenen esanguran duen eragina neurtzearren. Zehazki, $m = 2, 5, 10$ balioak erabili dira. Aurretik adibide batekin azaldu bezala, topiko kopuru ezberdin bakoitzerako LDA ereduak datu-sortako x_i mezu bakoitzaren errepresentazioa sortzen du, m-dimentsiodun espazio errealeko topiko-bektore baten bidez: $\mathbf{x}_i = (t_{i1}, t_{i2}, \cdots, t_{im}) \in \mathbb{R}^m$.

Errepresentazio honetarako **oinarri-lerro** izango den sailkatzaile gisa, guk proposatutako sistema bat erabili da. Honetan, $\mathbf{x}_i$ mezu bakoitzaren topiko-errepresentazioan gailentzen den topikoari ematen zaio garrantzia. Topiko hau, $tGailendu(\mathbf{x}_i)$ bezala adierazi da (1) adierazpenean. Izan bedi $\mathcal{X}(t_j, c_k)$ entrenamenduko instantzien azpi-multzoa halakoa non denetan gailentzen den topikoa t_j den eta gainera c_k klasekoak diren. Izan bedi $tcModa(t_j)$ t_j topiko nabarmenena duten instantzien arteko modako klasea (maizen agertzen den etiketa), (3) 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 azpi-multzoko 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) adierazpenean azaltzen den moduan. Hitz gutxitan esanda, instantziaren errepresentazioan gailentzen den topikoarentzat entrenamendu fasean gordetako modako klasea esleitzen zaio adibide berriari.

Aipatzekoa da $m = 5$ kasurako oinarri-lerroa eraikitzean, 5. topikoa gailendu den mezuen arteko modako

klasea suizidioarekin erlazionatutakoa izan dela, eta gainontzeko laurena aldiz, suizidioarekin erlazionatuta ez dagoena. Gure ustez, hau ongi islatzen da 1 irudiko hitz-hodeietan.

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

$$\mathcal{X}(t_j, c_k) = \{(\mathbf{x}_i; c_i) \in X_{Entrenamendua} | tGailendu(\mathbf{x}_i) = t_j \wedge c_i = c_k\} \tag{2}$$

$$tcModa(t_j) = \arg\max_{c_k \in \mathcal{C}} |\mathcal{X}(t_j, c_k)| \tag{3}$$

$$\widehat{c}(\mathbf{x}_{berria}) = tcModa(tGailendu(\mathbf{x}_{berria})) \tag{4}$$

Analizatutako m topiko kopuru bakoitzeko oinarri-lerroarekin lortutako emaitzak 3 taulan aurki daitezke. Asmatze-tasa guztiak %80tik gorakoak izan dira. Metrika honetan gailendu den topiko kopurua $m = 5$ izan da, estalduran eta F puntuazioan $m = 2$ nagusitu den bitartean.

3. Taula: **LDA errepresentazioen gainean oinarri-lerroko sailkatzailearekin** lortutako emaitzak (%)

m	Asmatze-tasa	Doitasuna	Estaldura	F puntuazioa
2	83,1	71,0	95,5	81,5
5	83,3	75,7	84,2	79,7
10	80,4	70,6	84,9	77,1

4 Ikerketaren ondorioak

Sare sozialetako mezuek suizidio zantzuak erakusten dituzten aztertzeko, ingelesez idatzitako Reddit sare sozialeko mezu sorta bat erabili da eta honen gainean sailkapen bitarreko ataza bat burutu da. Horretarako, datuen prestaketa sakona egin da lehendabizi. Ondoren, metodo ezberdinak erabili dira, gainbegiratuak zein gainbegiratu gabeak sailkapen bitarra egiteko.

Artearen egoerako metodo gainbegiratuekin eta mezuen errepresentaziorako hitz-embeddingak erabilita emaitza oso onak lortu dira, %97,9ko asmatze-tasara iritsiz. Aztertutako bost ereduen artean transformerrak izan dira portaerarik onena erakutsi dutenak, CNN eta LSTM sare-neuronalen aldean.

Bestalde, LDA era ez-gainbegiratuan erabiliz, mezuen zenbakizko errepresentazio ezberdin bat lortu dugu. Gainera, errepresentazio honen gainean burututako sailkapen atazetarako oinarri-lerroa proposatu dugu. Frogatu da ereduaren sinpletasuna kontuan izanik, topiko kopuru txikiko errepresentazioek emaitza onargarriak ematen dituztela (%83,3ko asmatze-tasa $m = 5$ izanik), eta ondorioz, errepresentazio horiek klaseekiko erlazioa dutela.

Nahiz eta eredu gainbegiratuekin lortutako emaitzak hobeak izan, hauek entrenatzeak dakarren kostu konputazionala nabarmenki handiagoa da dimentsionalitate baxuko LDA eredua eraikitzeak suposatzen duena baino. Gainera, topiko-ereduek eskainitako errepresentazioak askoz ere interpretagarriagoak dira, bektoreko posizio bakoitza topiko batekin erlazionatua baitago.

Arrazoi hauengatik, topiko-ereduak, ikerketa honetan landutako LDA eredua zehazki, sare sozialetan suizidio ideiagintza identifikatzeko lagungarriak izan daitezke, honek sortutako errepresentazio numerikoek sailkapen-klaseekiko erlazioa dutela frogatu baita.

5 Etorkizunerako proposatutako norabideak

Etorkizunerako lan gisa, alde batetik, datuen aurreprozesaketan emojiak tratatzea proposatzen da. Izan ere, karaktere berezi hauek emozioak adierazteko erabili ohi dira eta ataza honetan lagungarriak izan daitezke.

Bestalde, LDA erabiliz sortutako topiko-ereduak mezuen errepresentazio gisa erabilgarriak direla frogatu da. Gainera, topiko kopuru handiagoek mezuetatik informazio baliagarri gehiago erauztea ahalbidetuko dutela uste da. Sailkapenari dagokionean, oinarri-lerrotik haratago joan eta multzokatze-algoritmoak (*clustering*) erabil daitezke sailkapen bitarra burutzeko, edota ikasketa sakoneko sarrera gisa erabili hitz-embeddingekin konbinatuta.

Erreferentziak

Abdulsalam, A. & Alhothali, A. (2022). Suicidal ideation detection on social media: A review of machine learning methods. *arXiv preprint arXiv:2201.10515*.

Ananthakrishnan, G., Jayaraman, A. K., Trueman, T. E., Mitra, S., Abinesh, A., & Murugappan, A. (2022). Suicidal intention detection in tweets using bert-based transformers. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 322–327. IEEE.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blei David, M. & John, D. L. (2009). Topic models. *Text Mining*, 101–124.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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)*, 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Fodeh, S., Li, T., Menczynski, K., Burgette, T., Harris, A., Ilita, G., Rao, S., Gemmell, J., & Raicu, D. (2019). Using machine learning algorithms to detect suicide risk factors on Twitter. In *2019 International Conference on Data Mining Workshops (ICDMW)*, 941–948. IEEE.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for Latent Dirichlet Allocation. *advances in neural information processing systems*, 23.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through twitter in the us. *Crisis*.

Jurafsky, D. & Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Kolliakou, A., Bakolis, I., Chandran, D., Derczynski, L., Werbeloff, N., Osborn, D. P., Bontcheva, K., & Stewart, R. (2020). Mental health-related conversations on social media and crisis episodes: a time-series regression analysis. *Scientific reports*, 10(1):1–7.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mo, Y., Kontonatsios, G., & Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. *Systematic reviews*, 4(1):1–12.

Vargas, H. B. & Saavedra, J. E. (2012). Factores asociados con la conducta suicida en adolescentes. *Revista de Neuro-psiquiatría*, 75(1):19–19.

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). 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*, 24–33.

6 Eskerrak eta oharrak

Lan honetarako finantziazioa honako iturrietatik lortu da: Espainiako Gobernuko Zientzia eta Berrikuntzako Ministeritza (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31 eta TED2021-130398B-C22 MCIN/AEI /10.13039/501100011033 eta Europar Batasuneko NextGenerationEU/ PRTR funtsak) eta Eusko Jaurlaritzako diru-laguntzetatik (IXA IT1570-22 eta Sara Gracia ikasleari emandako diru-laguntza, 2022/07/11 EHAAko deialdiko erresoluzioan argitaratua).