VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

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

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Kontra-Narratiben Sorkuntza Automatikoa: Gorrotoa Deseraikiz

Irune Zubiaga, Jaione Bengoetxea, Aitor Soroa eta *Rodrigo Agerri*

157-164 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.19

Kontra-Narratiben Sorkuntza Automatikoa: Gorrotoa Deseraikiz

Irune Zubiaga, Jaione Bengoetxea, Aitor Soroa, Rodrigo Agerri

HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU. Manuel Lardizabal pasealekua, 1, 20018 Donostia-San Sebastian, Gipuzkoa

irune.zubiaga@ehu.eus

Laburpena

Lan honek ML-MTCONAN-KN aurkezten du, ezagutzan oinarritutako kontra-argudiaketarako lehen corpus eleaniztuna. Corpusa modu paraleloan egituratuta dago lau hizkuntzatan: ingelesa, euskara, gaztelania eta italiera. Datu-multzo honek ekarpena egiten du arloan, kontra-argudiaketarako baliabideen urritasunari aurre eginez, ezagutzan oinarritutako kontra-narratiben lehen bilduma eleaniztuna aurkeztuz. Gainera, euskaraz kontra-narratiba sorkuntza automatikorako teknika ezberdinen azterketa eskaintzen dugu, orain arte egin ez den ikerketa bat. Gure ekarpenaren helburua gorroto-diskurtsoei erantzun eraikitzaileak sortzeko gai diren sistemen garapena hobetzea da. ***Eduki-oharra:*** *Artikulu honek gorroto-diskurtsoa dauka, eta irakurle batzuentzat iraingarria izan daiteke. Ikuspegi horiek ez dute egileen pentsaera islatzen.*

Hitz gakoak: Adimen Artifiziala, Hizkuntzaren Prozesamendua, Argudiaketa

Abstract

This work introduces ML-MTCONAN-KN, the first knowledge-based multilingual corpus for counter-argumentation. The corpus is structured in a parallel manner across four languages: English, Basque, Spanish, and Italian. This dataset represents a significant advancement in the field, addressing the scarcity of resources for counter-argumentation and marking the first multilingual collection of knowledge-based counter-narratives. Additionally, we provide an in-depth analysis of automatic counter-narrative generation in Basque, a study that has not been conducted before. Our contribution aims to enhance the development of systems capable of generating constructive responses to hate speech. ***Content Warning:*** *This paper contains hate speech that may be offensive to some readers. These views do not reflect those of the authors.*

Keywords: *Artificial Intelligence, Natural Language Processing, Argumentation*

1 Sarrera eta motibazioa

Gaur egun ematen ari den gorroto-diskurtsoaren (GD) ugaritzeak, espazio digital eta fisikoetan, arrisku handiak dakartza kohesio sozialerako eta banakoen ongizaterako. Sare sozialen eta lineako foroen hedapenarekin batera, mezu iraingarriak eta estereotipo kaltegarriak zabaltzeko aukerak ere handitu egin dira. Gorroto-diskurtsoak bazterkeria eta polarizazioa areagotzen ditu, eta, ondorioz, gizarte demokratiko baten oinarriak ahultzeko arriskua dakar. Horregatik, funtsezkoa da gorroto-diskurtsoari aurre egiteko estrategiak garatzea, bai ikuspegi juridikotik, bai ikuspegi teknologiko eta pedagogikotik.

Tradizionalki, gorroto diskurtsoaren kontra erabili izan den estrategia errepresioan oinarritu da: edukiak ezabatzea eta hauen sortzaileak blokeatzea izan dira ohiko neurriak. Hala ere, neurri hauen eraginkortasuna zalantzan jarri da, hainbat muga baitituzte. Alde batetik, gorrotoa zabaltzeagatik blokeatutako erabiltzaileek diskurtso kaltegarriak zabaltzeko bide berriak aurkitu ohi dituzte. Bestetik, edukia ezabatzeak zentsuraren pertzepzioa piztu dezake erabiltzaileen artean, adierazpen-askatasunaren inguruko eztabaidak sortuz. Oraintsu, neurri hauen alternatiba gisa, kontra-narratibak (KN) sortzea estrategia itxaropentsu gisa agertu da. Kontra-narratiben helburua gorroto-diskurtsoa osatzen duten gezur eta zehaztasunik gabeko argudioak desegitea da, ikuspegi alternatibo eta eraikitzailea eskainiz. Diskurtso negatiboa ezabatu beharrean, enpatia eta ulermena sustatzea dute helburu, erabiltzaileen arteko elkarrizketa osasuntsua bultzatuz. Horrela, kontra-narratiben erabilerak lineako ingurune inklusiboagoa izatea ahalbidetzen du (Benesch, 2014; Schieb eta Preuss, 2016).

Metodo honek eraginkortasun potentzial handia duen arren (Hangartner *et al.*, 2021), kontra-narratibak sortzea lan zaila eta neketsua da. Gainera, gaur egun espazio digitaletan gorrotoa neurriz kanpo hedatzen ari da, eta horrek are zailago egiten du gorroto mezu bakoitzari kontra-narratiba batekin erantzutea. Arazo hori arintzeko, Hizkuntzaren Prozesamendua kontra-narratiben sorkuntza automatizatzeko bideak ikertzen ari da. Helburua erantzun hauek eskuz sortzen dituzten gobernuz kanpoko erakundeak (GKE) laguntzea da, haien lana erraztu eta prozesua eraginkorrago bihurtuko duten tresnez hornituz (Chung *et al.*, 2021c; Bonaldi *et al.*, 2024).

Kontra-narratiba sortzaile automatikoak garatu eta ebaluatzeko, ezinbestekoa da gorroto mezuak eta hauei aurre egiteko erabili diren kontra-narratiben adibideak (GD-KN pareak, alegia) biltzen dituzten datu-multzoak izatea. Lan honetan, horrelako datu-multzo bat aurkeztuko dugu: KN-MT-CONAN. Multzo honek adibide paraleloak eskaintzen ditu hainbat hizkuntzatan, hau da, datu-multzoko instantzia bera ingelesez, gaztelaniaz, italieraz eta euskaraz dago eskuragarri. KN-MT-CONAN aurkezteaz gain, lan honetan datu-multzoa erabiliz euskaraz dauden hainbat kontra-narratiba sortzaile automatiko ebaluatuko ditugu, eta sistema hauek erabilitako teknikak aztertu. Honen bidez, eremuaren egungo egoerari buruzko ikuspegi argia eskaini nahi dugu, eta metodo eraginkorrenak identifikatu. Bi ekarpen hauekin, kontra-narratiba sorkuntza automatikoaren oinarriak sendotzea dugu helburu, etorkizuneko garapenetarako abiapuntu sendo bat eskainiz.

2 Arloko egoera eta ikerketaren helburuak

Hizkuntzaren prozesamenduaren arloak urteak daramatza gorrotoaren aurkako borrokan lan egiten, gehien aztertu den arloa gorroto-diskurtsoaren detekzioa izan delarik (Alkomah eta Ma, 2022). Hala ere, kontra-narratiben sorkuntza automatikoa oraindik eremu berria da; izan ere, GD-KN pareez osatutako lehen corpusa 2019an sortu zen. CONAN (Chung *et al.*, 2019) izeneko corpus honek, musulmanei zuzendutako 6648 GD-KN pare biltzen ditu. Jatorriz italieraz, frantsesez eta ingelesez garatua, euskarara eta gaztelaniara ere itzuli zen (Bengoetxea *et al.*, 2024). Beranduago, 2021 urtean, MT-CONAN (Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco, 2021) sortu zen, 5003 GD-KN parez osatua. Islamofobia ez ezik, MT-CONANek beste 7 talderi zuzendutako gorroto-diskurtsoa jasotzen du, hala nola, emakumeak, LGBT+ komunitatea edo etorkinak. MT-CONAN gaztelerara itzuli da (Vallecillo-Rodríguez *et al.*, 2024), eta Korearrez GD-KN pareen corpus bat sortzeko oinarria izan da (Lee *et al.*, 2024).

Corpus hauek kontra-narratiba sorkuntza automatikorako metodoak eta sistemak garatzeko aukera eman dute. Sistema hauek, oro har, ikasketa sakonean eta hizkuntza-ereduetan oinarritzen dira, eta gehienek ingelesarekin egiten dute lan (Tekiroglu *et al.*, 2022; Halim *et al.*, 2023; Mathew *et al.*, 2018). Hizkuntza-ereduetan oinarritutako testu-sorkuntzako beste hainbat atazatan bezala, kontra-narratiba sorkuntza-sistemek informazio okerra sortzeko joera dute, haluzinazio izenez ezagutzen den fenomenoa (Perković *et al.*, 2024). Arazo honi aurre egiteko, ezagutzan oinarritutako kontra-narratiben sorkuntza proposatu da. Bertan, sistemari hainbat datu sendo eskaintzen zaizkio, gorroto-diskurtsoarekin batera, sorkuntza modu egokian bideratzeko (Chung *et al.*, 2021a; Jiang *et al.*, 2023).

Lan honetan, ML-MTCONAN-KN aurkezten dugu, ezagutzan oinarritutako GD-KN pareen datu-multzoa, zeina lau hizkuntzatan garatu den: ingelesez, euskaraz, italieraz eta gazteleraz. Corpusarekin batera, hainbat sorkuntza-sistema aurkeztuko ditugu eta haien arteko konparaketa egingo dugu, analisia euskarazko kontra-narratiben sorkuntzan zentratuz. Gure lanak ingurune eleaniztun batean sistema automatizatuak garatzeko baliabideak eskaintzea eta etorkizunean aplikazio praktikoak garatzeko gidalerroak ematea du helburu.

3 Ikerketaren muina

Lan honek bi ekarpen nagusi ditu: alde batetik, ezagutzan oinarritutako kontra-narratiben datu multzo eleaniztun berria aurkezten du; bestetik, euskaraz kontra-narratibaren sorkuntza automatikorako tekniken analisia aurkezten du, ataza ebazteko praktika onenak identifikatuz. Datu multzoaren deskribapena eta haren sorkuntzaren azalpena 3.1 atalean egingo da, eta kontra-narratiba sorkuntza sistemen deskribapena eta emaitzen analisia 3.2 atalean.

3.1 ML-MTCONAN-KN datu multzoa

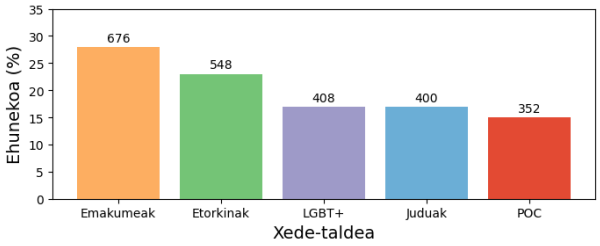
Atal honetan ML-MTCONAN-KN aurkezten dugu, ezagutzan oinarritutako kontra-narratiben sorkuntzarako corpus eleaniztuna. Instantzia bakoitza hiru elementuz osatuta dago: gorroto mezu bat (GD), mezu iraingarriarekin lotutako 5 ezagutza esaldi (EZ) eta kontra-narratiba bat (KN). Emandako 5 EZ esaldietako batzuk gorroto mezuaren aurka egiteko lagungarriak dira, informazio esanguratsua baitaramate, baina beste batzuk ez. 1. taulan

instantzia baten adibidea ikus daiteke, ingelesez eta euskaraz. Datu-multzoa jatorriz ingelesez garatu da, gero beste hizkuntzetara itzuliz. Jarraian, hirukoteetako elementu bakoitza nola lortu den deskribatzen da:

1. Taula: Corpuseko instantzia bat ingelesez (jatorrizko hizkuntza), eta euskaraz (itzulpena). Kontra-narratiba sortzeko erabili den EZ berdez nabarmendu da.

GD	We are caged in this world that is run by a secret elite of Jews and they rule it with an iron fist. Juduen elite sekretu batek zuzentzen duen mundu honetan kaiolaturik gaude eta burdinazko ukabilaz gobernatzen dute.
EZ	A country with roughly the population of Brooklyn was ruled by the iron fist of the ultraconservative Catholic Church. < EOS > Thousands of victims of clergy rape were revealed in the Catholic child-care network of orphanages, reformatories, hostels and seminaries. < EOS > In the event, his fabled iron fist fell far short of delivering a knock-out punch. < EOS > At Arafat, the pilgrims slit the throats of hundreds of thousands of sheep, cows and camels. < EOS > ÂAl-Maliki had earlier said he would crack down with an ”iron fist”on gangs in Basra. < EOS > Brooklyngo biztanleak zituen herrialde bat Eliza Katoliko ultrakontserbadorearen burdinazko ukabilak gobernatzen zuen. < EOS > Elizgizonen bortxaketaren milaka biktima agertu ziren umezurztegi, erreformatorio, ostatu eta mintegietako haurtzaindegi katolikoen sarean. < EOS > Hala eta guztiz ere, haren burdinazko ukabilak ez zuen kolpe bat ere eman. < EOS > Arafaten, erromesek lepoa moztu zieten ehunka mila ardi, behi eta gameluri. < EOS > ÂAl-Malikik lehenago esan zuen ”burdinazko ukabil”batekin jaitsiko zela Basorako banden gainean. < EOS >
KN	A country with roughly the population of Brooklyn was ruled by the iron fist of the ultraconservative Catholic Church. What evidence do you have that this is happening at a global scale because of Jews? Ia-ia Brooklyngo biztanleria kopurua duen herrialde bat Eliza Katoliko ultrakontserbadorearen burdinazko ukabilak gobernatu zuen. Zer froga duzu hau mundu mailan gertatzen ari dela esateko, juduen erruagatik?.

1. Irudia: Xede-taldeen araberako instantzien banaketa, kasu kopurua eta dagozkien ehunekoak erakutsiz.



- **GD:** Gorroto mezuak MT-CONANetik hartu dira. Bost talderi zuzendutako GDak bildu dira: juduak, LGBT+ komunitateko pertsonak, etorkinak, koloreko pertsonak, eta emakumeak. 1. irudian talde bakoitzari zuzendutako instantzia kopurua ageri da.
- **EZ:** Ezagutza hautatzeko Chung *et al.* autoreek (2021b) lanean deskribatzen duten ezagutza berreskuratze eta hautaketa modulua erabili da. Prozesuaren lehenengo pausoa GD bakoitzarentzat kontsulta bat sortzea da, ondoren kanpoko biltegi batean gorroto mezuarekin lotutako ezagutza adierazgarria bilatzeko. Bilaketak *Newsroom* (Grusky *et al.*, 2018) eta *WikiText-103* (Merity *et al.*, 2022) biltegietan egin dira *Solr* [1] bilaketatresna erabiliz. Azkenik, eskuratutako dokumentuak esaldietan zatitu dira eta ROUGE-L (Lin, 2004a) metrikaren arabera garrantzitsuenak diren 5ak hautatu dira EZ gisa. Corpusean ez dago zehaztuta ze jakintza den garrantzitsua eta zein ez.
- **KN:** Kontra-narratibak eskuz sortu dira EZn azaltzen den informazioan oinarrituta eta hurrengo anotazio-jarraibideak kontuan hartuta: (i) EZn ezagutza garrantzitsurik ez badago, adibidea baztertu, (ii) ezagutza garrantzitsua badago, erabili KNa idazteko, beharrezkoa bada egokituz eta (iii), EZ esaldi bat baino gehiago garrantzitsutzat jotzen bada, GD berarentzat bi KN desberdin idatzi daitezke, EZ-ren zati desberdinak erabiliz.

[1] `https://lucene.apache.org/solr/`

Prozesu honen bidez 2384 instantzia sortu dira ingelesez. Instantzia hauek honela banatu dira: 1584 trebakuntzarako, 400 balidaziorako eta 400 ebaluaziorako. Ondoren, instantzia bakoitza beste 3 hizkuntzara itzuli da: euskarara, italiarrera eta gaztelerara. Itzulpena bi urratsetan egin da. Hasteko datu-multzoa automatikoki itzuli da, DeepL[2] erabiliz gaztelaniarako eta italierarako eta Itzuli[3] euskararako. Ondoren, hizkuntza bakoitzeko hiztun natiboek automatikoki itzulitako GD eta KNen eskuzko berrikuspena egin dute, oker zeuden esaldiak zuzenduz.

3.2 Euskarazko kontra-narratiba sorkuntza sistemak

Atal honetan, ML-MTCONAN-KN datu multzoaz baliatuz, hizkuntza-ereduetan oinarritutako euskarazko kontra-narratiba sorkuntza sistema desberdinen eraginkortasuna ebaluatuko dugu. Nahiz eta sistema guztiek oinarritzat hizkuntza-ereduak izan, elkarren artean ezaugarri bereizgarriak dituzte. Sistemen propietate nagusiak 2. taulan azalatzen dira. Bertan ikus daitekeen moduan, zenbait sistema ML-MTCONAN-KNren trebakuntza zatia erabiliz doitu dira, baina beste batzuk ez dira atazarako doitu. Trebakuntza espezifikorik gabe atazak burutzea *zero-shot* izenez ezagutzen da (Brown *et al.*, 2020). Aldi berean, sistema batzuk euskaraz bakarrik trebatu dira beste batzuk 4 hizkuntzetako datuak erabili dituzten bitartean. Bigarren doiketa mota honi *doiketa eleaniztuna* deitzen diogu. Azkenik, sistema batzuek EZ bahetzeko teknikak erabili dituzte ezagutza garrantzitsua identifikatzeko, beste sistema batzuek EZ osoa kontuan hartuz edo EZ erabili gabe funtzionatzen duten bitartean. Izen bera baina identifikatzaile desberdina (adib., run1, run2, run3) dituzten sistemek, hurbilpen berberaren aldaerak dira. Aldaera hauek sistema konfigurazio desberdinak adierazten dituzte, eta konfigurazio bakoitzak errendimendua modu desberdinean optimizatzeko helburua du.

2. Taula: Euskarazko kontra-narratiba sorkuntza sistemak eta haien ezaugarriak hiru dimentsioren arabera: ea eredua doitu den, eredua hizkuntza anitzeko datuekin edo hizkuntza bakarreko datuekin doitu den eta ezagutza bahetu den ala ez.

Sistema	Doituta	Doiketa eleaniztuna	EZ baheketa
Hyderabadi Pearls (Farhan, 2025)	✓	-	-
TrenTeam (run 1) (Russo, 2025)	✓	✓	✓
Northeastern Uni (Wadhwa *et al.*, 2025)	✓	✓	-
NLP@IIMAS (run 1) (Preciado Márquez *et al.*, 2025)	✓	✓	-
TrenTeam (run 2-3)	-	-	✓
MilaNLP (Moscato *et al.*, 2025)	-	-	✓
NLP@IIMAS (run 2-3)	-	-	-
CODEOFCONDUCT (Bennie *et al.*, 2025)	-	-	-
HuaweiTSC (Lyu *et al.*, 2025)	-	-	-

Sistemak ebaluatzeko, kontra-narratiben sorkuntza automatikoan erabili izan ohi diren metrika tradizionalak erabili ditugu, hala nola BLEU (Papineni *et al.*, 2002), ROUGE-L (Lin, 2004b) eta BERTscore (Zhang *et al.*, 2020) antzekotasun metrikak, eta kreatibitatea neurtzen duen *Novelty* (Wang eta Wan, 2018) metrika. BLEU eta ROUGE-L metrikek sortutako testua eta erreferentziazkoa alderatzen dituzte n-gramen gainjartzea neurtuz. BERTscorek, aldiz, aurre-entrenatutako BERT ereduen bidez erreferentziaren eta sortutako testuaren errepresentazio bektorialak sortzen ditu eta haien arteko antzekotasun semantikoa neurtzen du. Azkenik, Novelty metrikak sortutako testuak entrenamendurako erabili diren datuetatik duen desberdintasun maila neurtzen du, sortutako edukia originalagoa edo errepikakorragoa den ebaluatzeko. Horrez gain, Zubiaga *et al.* autoreek (2024) lanean deskribatutako LLM bidezko ebaluazioa erabili da[4]. Sistemen *ranking*a azken metodo honekin egin da, giza-ebaluazioarekin metrika tradizionalek baino korrelazio handiagoa duela ikusi baita (Zubiaga *et al.*, 2024). Emaitzak 3. taulan aurkezten dira.

Taula erreparatuz ikusten da errendimendu onena duen sistema *CODEOFCONDUCT run1* dela; EZ erabiltzen ez duen doitu gabeko sistema. Hala ere, sorkuntza ataza asko ebaluatzeko egin izan den bezala BERTScore balioari

[2] https://www.deepl.com

[3] https://www.euskadi.eus/itzuli/

[4] Ebaluazio kodea https://github.com/hitz-zentroa/eval-MCG-COLING-2025 biltegian aurkitu daiteke.

3. Taula: Emaitza taula.

Pos.	Sistema	LLM Kalifikazioa	Metrika tradizionalak(%)				Sorkuntzaren Luzera
			ROUGE-L	BLEU	BERTscore	Novelty	
1	CODEOFCONDUCT run1	**1985.5**	8.2	1.5	66.4	86.8	67.5
2	CODEOFCONDUCT run3	1909.5	10.4	2.2	67.5	87.5	69.1
3	CODEOFCONDUCT run2	1896.0	9.8	2.2	67.0	87.1	66.2
4	NLP@IIMAS run2	1661.5	8.9	0.6	67.7	87.5	34.6
5	HuaweiTSC run2	1448.5	17.7	5.6	72.4	86.8	34.5
6	HuaweiTSC run3	1299.5	23.3	10.5	74.2	86.5	32.1
7	ground truth	1118.0	100.0	100.0	100.0	85.3	26.5
8	HuaweiTSC run1	1090.0	18.3	6.3	72.1	87.2	30.2
9	MilaNLP run1	1030.5	10.7	1.0	69.0	87.8	44.6
10	TrenTeam run2	1023.0	32.8	20.9	77.1	85.7	27.5
11	TrenTeam run1	1006.0	33.8	22.4	**77.6**	85.2	28.2
12	Hyderabadi Pearls run2	923.0	27.6	15.5	75.5	85.3	27.8
13	TrenTeam run3	917.0	31.7	18.2	76.6	85.9	24.0
14	Northeastern Uni run2	830.5	27.6	13.5	75.7	83.4	24.5
15	Northeastern Uni run3	827.5	30.9	17.6	76.2	85.2	29.6
16	Northeastern Uni run1	792.5	25.6	13.3	74.6	84.3	24.8
17	Hyderabadi Pearls run3	692.5	29.2	17.4	75.5	85.6	26.2
18	Hyderabadi Pearls run1	680.5	29.2	17.4	75.5	85.6	26.2
19	MilaNLP run3	545.5	18.5	6.9	70.4	87.4	50.5
20	MilaNLP run2	508.5	17.9	6.8	70.7	88.3	72.8
21	NLP@IIMAS run1	457.5	29.2	17.6	74.9	86.0	24.9
22	NLP@IIMAS run3	457.0	29.2	17.6	74.9	86.0	24.9

erreparatuko bagenio, *TrenTeam run1* izango litzateke irabazlea. Metriken arteko desadostasun hau aztertzeko sistema bakoitzak sortutako kontra-narratibetatik ausaz hautatutako 10 instantziaren eskuzko analisia egin da. 4. taulan instantzia hauetako bat aurkezten da. Bertan ikusten den moduan, LLMaren arabera irabazten duen sistemak ez du bere kontra-narratiba ezagutzan oinarritzen. Bitartean, BERTScoreren arabera irabazten duen sistemak EZ esaldiak hautatu eta hitzez hitz erabiltzen ditu. Horren ondorioz, antzekotasunean oinarritzen diren metriketan emaitza oso onak lortzen ditu, batzuetan zuzena ez den EZ ere erabili arren (4. taulan gorriz nabarmenduta).

Emaitzak era orokorrago batean aztertuz, doitutako sistemek antzekotasun metriketan balio altuagoak dituztela ikus daiteke. Horrez gain, sistema horiek sorkuntza-luzera laburragoak eta *Novelty* balio baxuagoak dituzte. Hau, doikuntzaren ondorioz, sistema horien irteera entrenamendurako erabili den datu-multzoaren distribuziotik hurbilago egotearen ondorioa da, ebaluazio multzoa distribuzio berberetik baitator. Bestalde, *zero-shot* teknikarekin sortutako kontra-narratibak luzeagoak dira eta LLM kalifikazio altuagoak lortzen dituzte. Honek lotura izan dezake LLM ebaluatzaileek emaitza luzeei puntuazio altuagoa emateko duten joerarekin (Hu *et al.*, 2024).

Orokorrean, argi ikusten da sistema hoberena aukeratzerakoan ebaluazio-metriken hautaketak duen eragina. LLM bidezko ebaluazioaren arabera, *zero-shot* teknika erabiltzen duten sistemak dira onenak, baina metrika tradizionalen arabera, BERTScore kasu, doitutako sistemak hobeak dira.

4 Ondorioak

Lan honek gorrotoaren aurkako borrokarako baliabide berri bat aurkezten du: lau hizkuntzatan paraleloki eraikitako ML-MTCONAN-KN kontra-narratiba corpusa. Bilduma hau bi mailatan da baliagarria: batetik, ezagutzan oinarritutako kontra-narratiba sorkuntzarako lehenengo datu-multzo eleaniztuna da, eta bestetik, euskara bezalako baliabide urriko hizkuntza bat barne hartzen du.

Horrez gain, gure ikerketak erakusten du gaur egun eskuragarri ditugun hizkuntza eredu handiak *zero-shot*en kontra-narratiba eraginkorrak sortzeko gai direla, askotan atazarako doituta dauden sistemak gaindituz. Honek, berriz, teknologia hauen potentziala adierazten du, baita baliabide urriko hizkuntzetan ere, gorrotoaren aurkako borrokan.

4. Taula: Euskarazko atazan ebaluazio partizioko GD instantzia bat, hari lotutako KNa eta hiru KN: lehena hizakiak idatzitakoa, bigarrena irabazlea LLMaren arabehera eta azkenik irabazlea BERTScore-F1 metrikaren arabehera. Erabilitako ezagutza zuzena berdez ageri da. Okerra berriz gorriz.

GD	Juduek Hollywood eta komunikabideak kontrolatzen dituzte.
EZ	< EOS > Horregatik, Hollywooden, judaismoa sotto voce izan zen beti. < EOS > Laikoak, eszeptikoak, ezkongabeak, anbibalenteak, desengainatuak, desengainatuak, batzuetan juduak: gaur egun, judaismoa eta juduen komunitatea dira nagusi. < EOS > Baina ia AEBetako zinema industriaren hastapenetatik, Hollywoodeko juduak asimilazioan aritu ziren, ez ospakizun erlijiosoan. < EOS > Eta eskatzen du Jainkoak berak zigor dezala juduen judaismoaren praktika oztopatzen saiatu zen errege oro. < EOS >
KN	Ia AEBetako zinema industriaren hastapenetatik, Hollywoodeko juduak asimilazioan zentratu ziren ziren, ez ospakizun erlijiosoan. Zergatik egon behar du lotuta haien erlijioa ustezko ”komunikabideen kontrol”batekin?. **KN_{LLM}** Judu komunitateak ez du Hollywood eta komunikabideak kontrolatzen, baina beren jarduerei eta ekintzak moralitate eta arau berriak ezartzen dizkiete, eta eztabaidagarriak diren edukiak saihesten ditu. Horrela, ez dago inolako frogarik juduak komunikabideak kontrolatzen ari direla, baizik eta beraien ekintzetan arau eta balio etikoak errespetatzen dituztela. **$KN_{B\text{-}F1}$** Baina ia AEBetako zinema industriaren hastapenetatik, Hollywoodeko juduak asimilazioan aritu ziren, ez ospakizun erlijiosoan. Gaur egun, judaismoa eta juduen komunitatea dira nagusi.

Kalifikazioa eta Arrazoiketa KN_{LLM} vs. $KN_{B\text{-}F1}$
8 6 Assistant 1’s response is more relevant and accurate to the user’s request. It provides a counter-narrative to the hate speech by explaining that the Jewish community does not control Hollywood and media, but they do influence the industry with their values and ethics. This response is helpful and detailed, hence the score of 8. Assistant 2’s response, while it does provide some historical context about the Jewish community’s involvement in Hollywood, it doesn’t directly address the hate speech. It doesn’t provide a counter-narrative to the hate speech, but rather provides a historical perspective. The response is somewhat relevant and accurate, but not as helpful or detailed as Assistant 1’s response, hence the score of 6

5 Etorkizunerako planteatzen den norabidea

Etorkizunean, corpusa hizkuntza gehiagotara zabaldu nahi da. Horrez gain, kontra-narratiben zuzentasun faktuala neurtzeko, EZ zuzena zein den anotatu nahi da eta ebaluazio metodo berriak aztertu.

Erreferentziak

Alkomah, Fatimah, eta Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information* 13.273.

Benesch, Susan. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876* .

Bengoetxea, Jaione, Yi-Ling Chung, Marco Guerini, eta Rodrigo Agerri. 2024. Basque and Spanish counter narrative generation: Data creation and evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2132–2141, Torino, Italia. ELRA and ICCL.

Bennie, Michael, Bushi Xiao, Chryseis Xinyi Liu, Demi Zhang, Jian Meng, eta Alayo Tripp. 2025. CODEOFCONDUCT at multilingual counterspeech generation: A context-aware model for robust counterspeech generation in low-resource languages. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 37–46, Abu Dhabi, UAE. Association for Computational Linguistics.

Bonaldi, Helena, Greta Damo, Nicolás Ocampo, Elena Cabrio, Serena Villata, eta Marco Guerini. 2024. Is safer better? the impact of guardrails on the argumentative strength of llms in hate speech countering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3446–3463.

——, María Estrella Vallecillo-Rodríguez, Irune Zubiaga, Arturo Montejo-Raez, Aitor Soroa, María-Teresa Martín-Valdivia, Marco Guerini, eta Rodrigo Agerri. 2025. The first workshop on multilingual counterspeech

generation at COLING 2025: Overview of the shared task. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 92–107, Abu Dhabi, UAE. Association for Computational Linguistics.

Brown, Tom, Benjamin Mann, Nick Ryder, eta et al. Subbiah. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901.

Chung, Yi-Ling, Marco Guerini, eta Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

——, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, eta Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2819–2829, Florence, Italy. Association for Computational Linguistics.

——, Serra Sinem Tekiroğlu, eta Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 899–914, Online. Association for Computational Linguistics.

——, Serra Sinem Tekiroğlu, Sara Tonelli, eta Marco Guerini. 2021c. Empowering NGOs in countering online hate messages. *Online Social Networks and Media* 24.100150.

Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Farhan, Md Shariq. 2025. Hyderabadi pearls at multilingual counterspeech generation : HALT : Hate speech alleviation using large language models and transformers. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 65–76, Abu Dhabi, UAE. Association for Computational Linguistics.

Grusky, Max, Mor Naaman, eta Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Halim, Sadaf MD, Saquib Irtiza, Yibo Hu, Latifur Khan, eta Bhavani Thuraisingham. 2023. Wokegpt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–10. IEEE.

Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, eta others, 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment.

Hu, Zhengyu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Jingang Wang, Zhenyu Chen, eta Hui Xiong. 2024. Rethinking llm-based preference evaluation. *arXiv preprint arXiv:2407.01085* .

Jiang, Shuyu, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, eta Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650* .

Lee, Seungyoon, Chanjun Park, DaHyun Jung, Hyeonseok Moon, Jaehyung Seo, Sugyeong Eo, eta Heuiseok Lim. 2024. Leveraging pre-existing resources for data-efficient counter-narrative generation in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10380–10392, Torino, Italia. ELRA and ICCL.

Lin, Chin-Yew. 2004a. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81, Barcelona, Spain. Association for Computational Linguistics.

——. 2004b. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Lyu, Xinglin, Haolin Wang, Min Zhang, eta Hao Yang. 2025. HW-TSC at multilingual counterspeech generation. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 47–55, Abu Dhabi, UAE. Association for Computational Linguistics.

Mathew, Binny, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, eta Animesh Mukherjee. 2018. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media*.

Merity, Stephen, Caiming Xiong, James Bradbury, eta Richard Socher. 2022. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Moscato, Emanuele, Arianna Muti, eta Debora Nozza. 2025. MNLP@multilingual counterspeech generation: Evaluating translation and background knowledge filtering. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 56–64, Abu Dhabi, UAE. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, eta Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Perković, Gabrijela, Antun Drobnjak, eta Ivica Botički. 2024. Hallucinations in llms: Understanding and addressing challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 2084–2088.

Preciado Márquez, David Salvador, Helena Gómez Adorno, Ilia Markov, eta Selene Baez Santamaria. 2025. NLP@IIMAS-CLTL at multilingual counterspeech generation: Combating hate speech using contextualized knowledge graph representations and LLMs. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 29–36, Abu Dhabi, UAE. Association for Computational Linguistics.

Russo, Daniel. 2025. TrenTeam at multilingual counterspeech generation: Multilingual passage re-ranking approaches for knowledge-driven counterspeech generation against hate. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 77–91, Abu Dhabi, UAE. Association for Computational Linguistics.

Schieb, Carla, eta Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, 1–23.

Tekiroglu, Serra, Helena Bonaldi, Margherita Fanton, eta Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3099–3114.

Vallecillo-Rodríguez, María-Estrella, Maria-Victoria Cantero-Romero, Isabel Cabrera-De-Castro, Arturo Montejo-Ráez, eta María-Teresa Martín-Valdivia. 2024. CONAN-MT-SP: A Spanish corpus for counter-narrative using GPT models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3677–3688, Torino, Italy. ELRA and ICCL.

Wadhwa, Sahil, Chengtian Xu, Haoming Chen, Aakash Mahalingam, Akankshya Kar, eta Divya Chaudhary. 2025. Northeastern uni at multilingual counterspeech generation: Enhancing counter speech generation with LLM alignment through direct preference optimization. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 19–28, Abu Dhabi, UAE. Association for Computational Linguistics.

Wang, Ke, eta Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 4446–4452. International Joint Conferences on Artificial Intelligence Organization.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, eta Yoav Artzi, 2020. Bertscore: Evaluating text generation with bert.

Zubiaga, Irune, Aitor Soroa, eta Rodrigo Agerri. 2024. A LLM-based ranking method for the evaluation of automatic counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 9572–9585, Miami, Florida, USA. Association for Computational Linguistics.

6 Eskerrak eta oharrak

Lan hau COLING 2025eko *The First Workshop on Multilingual Counterspeech Generation*en barne antolatutako MCG-COLING-2025 (Bonaldi *et al.*, 2025) ataza partekaturako egindako lanean oinarrituta dago.

Ikerketa Euskadiko Gobernuak (IT-1805-22 Ikerketa-talde finantzaketa) partzialki diruz lagunduta egin da. Era berean, eskerrak eman nahi dizkiegu MCIN/AEI/10.13039/501100011033-ko proiektu hauei: (i) DeepKnowledge (PID2021-127777OB-C21) eta FEDER, EU-ko; (ii) Disargue (TED2021-130810B-C21) eta Europar Batasunaren NextGenerationEU/PRTR-eko; (iii) DeepMinor (CNS2023-144375) eta (iv) DeepR3 (TED2021-130295B-C31)