Iruzurrezko portaeren detekzioa *crowd* motako etiketazioan

de Velasco Vázquez, Mikel; López Zorrilla, Asier eta Justo Blanco, Raquel

Euskal Herriko Unibertsitatea UPV/EHU

mikel.develasco@ehu.eus

Laburpena

Lan honek crowd motako etiketazioan agertu daitezkeen kalitate baxuko etiketak detektatzea du helburu. Proposatutako metodologia balioztatzeko, saiakuntzak ataza zail eta subjektibo batekin egin ditugu: emozioen detekzioarekin. Iruzurrezko langileak topatzeko zenbait neurri proposatu dira, etiketatze denboran, langileen arteko adostasunean eta langileen erantzunen banaketan oinarriturikoak. Neurri bakoitza baliagarria dela frogatu dugun arren, gure ondorio nagusia neurriak batzerakoan iruzurrezko langileak detektatzeko probabilitatea handitzen dela da.

Hitz gakoak: Metodo gainbegiratuak, Etiketazioa, Iruzurrezko jokaera, Jendetza, Subjektibotasuna

Abstract

This work aims at detecting low quality labels in crowdsourcing annotation tasks. We validate our proposal carrying out experiments in a difficult and subjective task: emotion recognition. We have developed several measures in order to detect fraudulent behaviour, including measures related to the labelling time, worker inter-agreement and the distribution of the answers. Not only do we show that each of the described measures is helpful but we also demonstrate that mixing them is the best way to go.

Keywords: Supervised learning, Annotation, Fraudulence behaviour, Crowdsourcing, Subjective

1. Sarrera eta motibazioa

Adimen artifizialaren barruan, metodo gainbegiratuek (*supervised methods*), etiketaturiko data kantitate handiak erabili ohi dituzte eredu sendoak sortzeko. Hala ere, etiketatze prozesua, askotan, esfortzu, denbora eta diruaren beharra duen lana da. Tradizionalki, lan espezifikorako heziak izan diren etiketatzaile adituak erabiltzen dira prozesu honetarako, baina halako etiketatzaileak aurkitzea nahiko zaila da, prozesu konplexua eta garestia bihurtuz. Gainera, ataza subjektiboetan, emozioen analisian adibidez, jende askoren iritzia etiketatzaile adituen iritzia baino interesgarriagoa izan daiteke.

Azken urteetan, *crowdsourcing*-a (jendetza bitartez sortutako etiketazioa) etiketazioak egiteko eraginkorragoa den alternatiba berria bilakatu da. Izatez, informazioaren berreskurapenean, hizkuntza naturalaren prozesamenduan, zein beste eremutara lotuta dauden arlo ezberdinetan erabili ohi da. Crowdsourcing plataformen artean, Amazoneko Mechanical Turk[1] edo SamaSource[2] aipa daitezke. Metodologia honek, etiketatze lana etiketatzaile anitz ezberdinekin burutzea ahalbideratzen du, etiketazio lan oso bat mikro-lan askotan zatituz. Gainera kasu askotan langile askoren iritzia kontuan hartzea onuragarria izan daiteke.

Lan honetan, oso subjektiboa den ataza batekin lan egin dugu, audio-hizketa segmentuetan emozioen detekzioan hain zuzen ere. Emozioak anbiguotasunik gabe definitu ezin direnez (Ortony eta Turner, 1990), jende anitzen iritizia oso interesgarria eta baliotsua da. Anbiguotasun horrek iritzien arteko desadostasuna sor dezake, entzulearen pertzepzioarengatik ala ingurune kulturalarengatik sortutakoa adibidez (Gurney, 1884; Scherer, 1999). Hori dela eta, etiketatze prozesua burutzeko metodorik aproposena crowdsourcing-ekin egitea da, ahal bezain beste iritzi kontuan hartzeko.

[1] www.mturk.com
[2] www.samasource.org

Dena den, ondoren azalduko diren arrazoiengatik, crowdsourcing-ak kalitate baxuko etiketak sor ditzake. Kalitate baxuko etiketen kausa nagusia langileen jarreraren ondorioa da. Horregatik, lan hau jarrera desegoki bat duten langileak detektatzera bideratuta dago. Iruzurrezko langilak topatzeko zenbait neurri proposatzen ditugu, etiketatze denboran, langileen arteko adostasunean eta langileen erantzunen banaketan oinarriturikoak.

2. Arloko egoera eta ikerketaren helburuak

Crowdsourcing-a nahiko hedatuta egon arren eta datu bilketa zein datu-etiketak balioztatzeko baliagarria izan arren, metodo honekin lortutako emaitzen kalitatea oraindik dudan dago. Izan ere, komunitateak arlo honetan ahalegin handiak egin arren (Eickhoff eta de Vries, 2011; Gennaro *et al.*, 2010; Gadiraju *et al.*, 2015), datuen fidagarritasuna neurtzea ez da bat ere erraza. Eickhoff autoreen (2013) lanaren arabera hainbat langile desegoki aurki daitezke: gutxiengo baldintzak betetzen ez dituzten langile ezgaituak alde batetik, erantzun okerrak emanez esperimentua baliogabetzen saiatzen diren langile maltzurrak bestetik, eta lanari beharrezko aditasuna jartzen ez dioten langile adigabeak azkenik.

Fidagarriak ez diren langileak antzemateko biderik arruntena *gold-standard* deritzon aurredefinitutako mikro-lanak mikro-lan arruntekin nahastea da. Era honetan, erantzun ezaguneko galdera bat oker erantzuten duten etiketatzaileak baztertzen dira. Metodo hau lan askotarako fidagarria izan arren, ez da baliozkoa kasu guztietan, galdera irekiak duten lanetarako, adibidez. Era berean, langile maltzurrak gold-standard lana sahiesteko era berriak bilatu ditzakete, kontu berrietan informazioa bererabiltzeko testen galderak ikasiz (Rothwell *et al.*, 2015) besteak beste. Gainera, ataza subjektiboetan, gold-standard metodoak ez dauka zentzu askorik, anbiguotasun gabeko kasuak aukeratzea, berez, ariketa subjektiboa delako. Adibidez gure ataza eta audioaren transkripzioa erabat desberdiak dira. Audioaren transkripzioan gertatzen ez den bezala, gurean ezin da objektiboki jakin zein den egiazko etiketa.

Gold-standard metodologia desegokia den kasuetarako beste irtenbide batzuk garatu dira. Adibidez, Filatova autoreek (2012) lanean ironia eta sarkasmoa etiketatzean, gehiengoaren botazioan oinarritutako algoritmo bat erabili zuten kalitatea kontrolatzeko (Ipeirotis *et al.*, 2010). Dena den, gehiengoaren botazioa erabiltzeko eta kalitate kontrol ona ziurtatzeko, mikro-lan bakoitzeko etiketa ugari behar dira eta horrek etiketazio lana garestitzen du. Ipeirotis *et al.* autoreen (2010) lana etiketatzaileen arteko adostasunean oinarrituta dago, eta Dawid eta Skene autoreen (1979) lanean inspiratuta dagoen itxaropen-maximizatze algoritmo bat erabiltzen du. Hala eta guztiz ere, etiketatzaileen arteko adostasunak huts egin dezake langile maltzurrak detektatzerakoan, sistema automatikoen bidez erantzunen atzean dagoen banaketa ikasi dezaketelako. Arazo honi irtenbidea emateko, Raquel Justo autoreek (2017) laneko errore-tasa kalkuluan oinarritzen den algoritmo bat proposatzen dute.

Etiketatzaileen arteko adostasun neurketen artean bateratzeko lanak egon arren (Klaus, 2011), neurri desberdinen artean nahasmen handia dago. Dakigunaren arabera, ez dago ataza subjektiboko datuen kalitatea neurtzeko fidagarria den metodorik. Hori dela eta, lan honetan zehar iruzurrezko langileen detekziorako hainbat neurri proposatzen ditugu eta euren konbinazio lineala erabat ereaginkorra dela erakusten dugu.

3. Iruzurrezko langileen detekzioa

Atal honetan, hasteko iruzurrezko langileak topatzera bideratuta dauden neurriak azaltzen ditugu. Geroago, neurri hauek guztiak zein testuingurutan balioztatu diren azalduko dugu. Azkenik, saiakuntzetan lortutako emaitzak erakusten ditugu.

3.1. Data-kalitatearen neurketa

Crowdsourcing teknikak erabiliz lortutako etiketazioen osteko analisia egiteko, denboran oinarritutako neurketa simpleetatik etiketatzaileak beraien artean duten adosmena neurtzen duten neurrietara arteko neurrriak aztertu genituen. Neurri hauek adierazteko, hau da lanean zehar erabiliko dugun notazioa:

- **Etiketatzaile bat** edo **langile bat** adierazteko, l indizea erabiliko da. Adb. n_l, l langileak egindako mikro-lanen kopurua izango zen.
- **Mikro-lan** bakoitza adierazteko, s indizea erabiliko da. Adb. n_s, mikro-lan bakoitzeko erantzunen kopurua izango zen. Gure ikerketaren barruan, segmentu bakoitza 5 aldiz erantzun da.
- **Galdera bakoitzaren erantzun posibleak** adierazteko, e indizea erabiliko da. Adb. n_e, e erantzun posiblea zenbat aldiz hautatu den.

3.1.1. Etiketatze denboran oinarritutako neurketak

Denborarekin erlazionatuta dauden bi informazio desberdin erabili ditugu. Alde batetik, langile bakoitzak etiketatzen eman duen gehiengo denbora (T_l), eta bestetik, mirko-lan bakoitzaren etiketa sortzeko behar izan duen denbora (t_e), hau da etiketa bakoitza sortzeko erabilitako denbora. Datu hauekin, langile bakoitzeko hiru neurketa desberdin kalkulatzen ditugu:

1. Lanaldian emandako denborarik luzeena (T_l). Neurri honek mikro-lanen artean 10 minutuko etenaldirik gabe lan egin duen gehienengo denbora adierazten du.
2. Mikro-lan bat burutzeko erabilitako bataz besteko denbora, (1) formulan azaltzen den bezala.

$$\overline{t_l} = \frac{\sum t_{ls}}{n_l} \tag{1}$$

3. Mikro-lan bat burutzeko erabilitako denboraren desbiderapen estandarra, (2) formulan azaltzen den bezala.

$$\sigma(t_l) = \sqrt{\frac{1}{n_l}\sum(t_{ls} - \overline{t_l})^2} \tag{2}$$

3.1.2. Etiketatzaileen arteko adostasunean oinarritutako neurketak

Crowdsourcing teknikak erabiliz ateratako datuen adostasuna kalkulatzeko hainbat lan egin dira, (Bennet *et al.*, 1954; Scott, 1955; Cohen, 1960; Krippendorff, 2004) adibidez. Neurri hauek etiketazio guztiak batera duten adostasuna neurtzen dute. Guk langileak deskribatzeko neurriak bilatzen ditugunez, adostasun neurri hauek ez dira zuzenenan aplikagarriak. Raquel Justo autoreen (2017) lana jarraituz langile mailako neurri bat definituko dugu adostatsun neurri batetik abiatuz.

Modu honetan, (3) ekuazioan agertzen den bezala langile mailako adostasun neurria kalkulatzen da, langile guztiekin kalkulatutako adostasuna (A) eta langile bat gabe kalkulatutako adostatuna (A_l) erabiliz.

$$\Delta A_l = \frac{A - A_l}{n_l} \tag{3}$$

Bi langileen arteko adostasuna neurtzeko modurik sinpleena ehuneko adostasuna ala antzemandako adostasuna da, (4) ekuazioan ikusten den bezala.

$$A_o = \frac{1}{n_I}\sum_{i \in I} agr(s_{l_1 i}, s_{l_2 i}) \tag{4}$$

non I bi langile etiketaturiko mikro-lanen arteko ebakidura ($S_{l1} \cap S_{l2}$) den; n_I bi langileak etiketaturiko mirkolan kopurua; eta $agr(s_{l_1 i}, s_{l_2 i})$, l_1 langileak i mikro-lanerako emandako emaitzaren eta l_2 langileak i mikro-lan berdinerako emandako emaitzaren adostasun funtzioaren emaitza.

Galdera ezberdinetarako, adostasun funtzio ezberdinak izan ditzakegu, erantzun posibleen artean dauden erlazioen arabera. Adibiderik errezena hurrengoa da:

$$agr_{i,k} = \begin{cases} 1 : \text{bi langileek etiketa bera jartzen badute} \\ 0 : \text{bi langileek etiketa ezberdina jartzen badute} \end{cases} \tag{5}$$

Hala ere, erantzun posibleak ordenaren bat jarraitzen badute, adostasun funtzio egokiago bat sor dezakegu, gehiengo adostasuna 1 eta gutxiengo adostasuna 0 izanik. Adibidez, hiru erantzun posibleak eskala batean irudikatu ahal badira, “Asko”, “Gutxi” eta “Ezer ez” esate baterako, funtzio egoki bat hurrengoa liteke (6):

$$agr_{i,k} = \begin{cases} 1 & : \text{bi langileak etiketa bera jartzen badute} \\ 0.5 & : \text{etiketa bakar bat “Gutxi” bada} \\ 0 & : \text{beste kasuetan} \end{cases} \quad (6)$$

Antzemandako adostatsuna literaturan aurkezten den adostasun neurri hedatuena izan arren (Artstein eta Poesio, 2008), ez du adostasunak ausaz gertatu daitezkeela kontuan hartzen. Hori dela eta, antzemandako adostasuna ausazko adostasunarekin egokitu ahal da, 7 ekuazioan ageri den Krippendorff autoreek (2004) lanean azalduriko α neurria bezala.

$$\alpha = \frac{A_o - A_e}{1 - A_e} \quad (7) \qquad A_e = \sum_{i \in I} P(i|l_1) \cdot P(i|l_2) \quad (8)$$

non A_e itxarondako ausazko adostasuna den. Itxarondako adostasuna, (8) ekuazioari jarraituz, ausaz erantzuten duten bi langileen artean (l_1 eta l_2) dagoen adostasuna neurtzen du, I bi langileek etiketaturiko mikro-lan berdinen multzoa izanik.

Krippendorff autoreen (2004) lanean ageri den moduan, α koefizientea (9) formularen arabera kalkulatu dezakegu. (10) formulan o_{e_c,e_k} balioak kalkulatzeko formuaren bidez, e_c eta e_k erantzun posibleen artean antzemandako desadostasuna kalkulatzen da. Bestalde, (11) formulan e_c erantzun posiblerako antzemandako desadostasuna kalkulatzen da. (9) formulan ikusten den agr_{e_c,e_k}-rako, (5) eta (6) adosmen funtzioak erabil daitezke, kasuaren arabera.

$$\alpha = 1 - (\sum_{c \in E} o_{e_c} - 1) * \sum_{c \in E} \sum_{k \in E} \frac{o_{e_c,e_k} * agr_{e_c,e_k}}{o_{e_c} * o_{e_k} * agr_{e_c,e_k}} \quad (9)$$

$$o_{e_c,e_k} = \sum_{s \in S} \frac{n_{e_c s} * n_{e_k s}}{n_s - 1} \quad (10) \qquad o_{e_c} = \sum_{k \in E} o_{e_c,e_k} \quad (11)$$

Krippendorff autoreek (2004) lanean azalduriko α nola kalkulatzen den jakin eta gero, aurretik aipatutako ΔA_l kalkulatzen dugu α balioarekin ($\Delta\alpha_l = \frac{\alpha - \alpha_l}{n_l}$). $\Delta\alpha_l$-ren balio negatiboak adostasunaren hobekuntza adierazten du. Beraz $\Delta\alpha_l$ zenbat eta txikiagoa izan orduan eta handiagoa izango da adostasuna.

$\Delta\alpha_l$-k langile baten eragina neurtzen du, hala ere, beste adostasun eta elementu batzuk baita neurrian eragina dute. Hori dela eta, Raquel Justo autoreek (2017) lanean proposaturiko β_l neurria erabili dugu ere. β_l neurriak langile baten eta mikro-lan berberak etiketaturko beste langileen arteko antzemandako adostasuna neurtzen du, (12) formulan agertzen den bezala.

$$\beta_l = \frac{1}{|S_l|} \sum_{s \in S_l} \sum_{m \in L_s - l} \frac{agr_{m,l}}{|L_s| - 1} \quad (12)$$

non S_l, l langileak egindako mikro-lan multzoa den; L_s, s mikro-lana egin duten langileak eta $agr_{m,l}$, (5) zein (6) adostasun funtzioa izan daiteke, kasuaren arabera.

3.1.3. Erantzunen banaketetan oinarritutako neurketak

Beste alde batetik, oso interesgarria izan daiteke, ausazko erantzunak zein erantzun bakarra erantzuten duten langileak detektatzea. Horretarako langile bakoitzak etiketaturiko datuen banaketa (b_l), langile guztiek batera etiketaturiko datuen banaketarekin (B) konparatzen duen metodo bat proposatzen dugu. Metodo honek banaketen distantzia neurtzen du (13).

$$d = \text{distantzia}(b_l, B) \quad (13)$$

Gure lanean bi distantzia mota desberdinekin lan egin dugu. d_1 distantzia euklidear arrunta da (14) eta d_2 distantziak probabilitate banaketan magnitude orden desberdineko elementuak egotea gehiago zigortzen du (15).

$$d_1(b_l, B) = \sum_{e \in E} (\mathbf{l}_e - \mathbf{B}_e)^2 \quad (14) \qquad d_2(b_l, B) = \sum_{e \in E} \left(\frac{\max(\mathbf{l}_e, \mathbf{B}_e)}{\min(\mathbf{l}_e, \mathbf{B}_e)} - 1 \right) \quad (15)$$

non b_l langilearen probabilitate banaketa den eta **B** langile gusztien probabilitate banaketa den, $\mathbf{l}_e$, eta $\mathbf{B}_e$ haien probabilitate banaketaren elementu bat izanik.

3.2. Ataza definizioa eta etiketatze prozesua

Aurreko atalean azaldutako neurri guztiak emozioak etiketatzeko lan batean aztertu nahi ditugu. Horretarako, crowdsourcing teknikak erabili dira "La Sexta Noche" corpusa emozioz etiketatzeko. "La Sexta Noche" telebista sailean emozio naturalak agertzen direnez, nahiko arraroa da mutur-emozioak aurkitzea, zalantzazko etiketak sortuz. Beste alde batetik, sei orduko telesaiak direnez, emozio bakarra aurkitu daitekeen tarteetan segmentatu behar dira. Segmentu horien luzera emozio bat adierazteko bezain luzeak eta emozioa aldaketa bat ez agertzeko bezain laburrak izan behar dute. Hortaz, "La Sexta Noche" corpusa 3 eta 5 segundu arteko segmentuetan banatu dugu, crowdsourcing-en bidez etiketa emozionala jartzeko.

Banatu dugun segmentu bakoitzaren emozioa adierazteko hainbat eredu aurki daitezke. Alde batetik emozioen eredu kategorikoa daukagu (Ekman, 1992), eta bestetik eredu dimentzionala (Gunes eta Pantic, 2010; Russell eta Mehrabian, 1977). "La Sexta Noche" telesailaren mintzagaien artean, politika, gertaera eta berriak zein eztabaidak maiz agertzen direla kontuan hartuz eta esperimentu txiki batzuk aurre-eginez, 10 emozio ezberdineko bilduma sortu genuen corpus hau etiketatzeko. Sortutako bildumarekin eredu kategoriko eta beste galdera batzuekin eredu dimentzionalako galdetegia sortu genuen:

1. Nola hantzematen duzu hizlaria?
   - Aztoratuta
   - Zertxobait aztoratuta
   - Neutrala
2. Bere gogo egoera hurrengoa da:
   - Positiboa (atsegina / eraikitzailea)
   - Zertxobait positiboa
   - Neutrala
   - Zertxobait negatiboa
   - Negatiboa (desatsegina / kolaboragaitza)
3. Nola ikusten duzu hizlaria egoera horretan?
   - Egoera kontrolatzen
   - Ez du egoera kontrolatzen ezta jarrera defentziboan egoten
   - Egoera defensiboan
4. Aukeratu hizlariaren emozioa hoberen deskribatzen duen aukera:
   - Lasaia / Axolagabe
   - Azpertuta / Nekatuta
   - Hunkituta
   - Harrituta
   - Alaia
   - Interesdun
   - Kezkatuta
   - Lotsatia
   - Haserre
   - Tentsoa

Galdetegia definitu eta gero, ikerketa taldearen baliabideen artean dagoen CrowdZienzia[3] crowdsorcing plataformarekin (Justo *et al.*, 2016) etiketatu dugu. Plataforma hau erabiltzeko arrazoien artean: 1) Amazoneko Mechanical Turk plataforma bakarrik Estatu Batuetan dagoela erabilgarri; 2) beste plataforma batzuek, SamaSource bezala, gaztelania mintzatzen dutenen artean %20 baino gutxiago espainarrak direla, eta ataza honetarako Espainian jaio diren pertsonak baliagarriagoak dira Amerikako gaztelainarekin dagoen diferentziagatik; eta 3) gure arteko etiketatze probak egiteko etiketatzaile finko batzuk edukitzea ahalbidetzen digula daude.

[3] Komunitate zientifikorako erabilgarri, muga zehatz batzuen artean. `https://crowdzientzia.ehu.eus`

3.3. Saiakuntza eta emaitzak

Aurreko ataletan azaldutako neurriak kontuan hartuta, CrowdZientzia plataformarekin sortutako datuak aztertuko ditugu. Datu hauek corpusetik auzaz aukeratutako 5.500 segmentuetatik datoz eta etiketak sondoak izateko segmentu bakoitza 5 aldiz etiketatzea nahikoa zela erabaki genuen, 27.500 mikro-lan sortuz. Mikro-lan guztiak egiteko 129 langilek parte hartu zuten, baina langile guztiek ez zuten lanaren proportzio bera etiketatu. Gutxien etiketatu zuen langileak mikro-lan bakarra etiketatu zuen, 2081 mikro-lan etiketatu zituen lan gehien egin zuen langileak, eta batez bestekoa 213 mikro-lan izan ziren. Langileei Amazoneko txekeekin ordaintzea erabaki genuen.

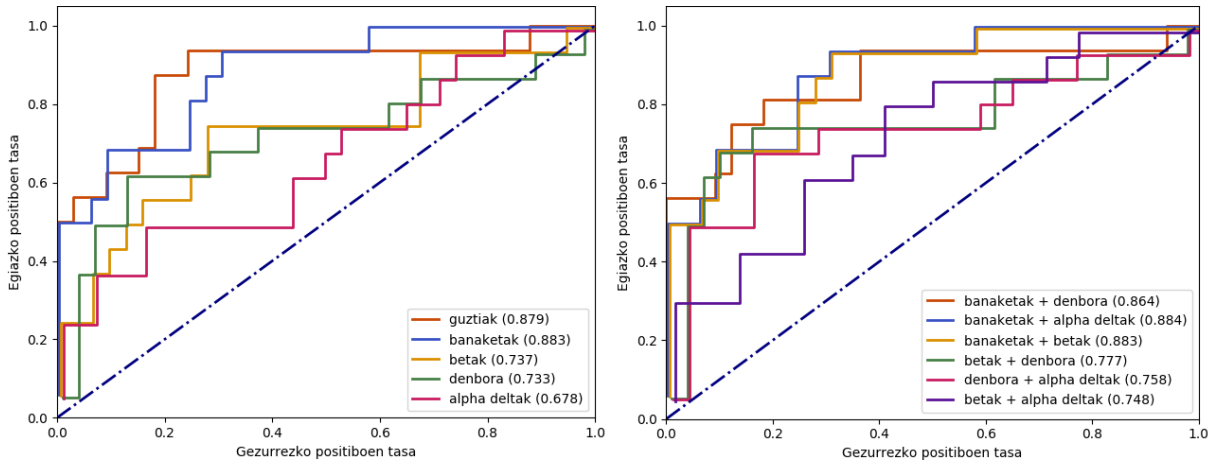
3.1 ataleko neurriak konparatzeko, aztertutako neurriren batean 20 langile txarren artean zeuden langileak hartu genituen (45 langile, langile gehienek neurri askotan txarrenetarikoen artean zeudelako) eta eskuz klasifikatu genituen iruzurrezko langile ala langile egoki bezala. Honetarako, aipatutako langile bakotzaren lana bi pertsonen artean aztertu eta iruzurrezko langilea ala langile egoki bezala sailkatzea eztabaidatu genuen. Ikerketa honetan, 6 iruzurrezko langile eta zalantzazko 10 iruzurrezko langile aurkitu genituen (3.2 atalean azaldutako galdetegiaren galdera batzuk txarto egiten zutela dirudielako).

Ondoren, aztertutako 45 langileak (6 iruzurrezko langile eta 39 langile egoki) zeinbait sailkatzaile linealekin doitu genituen. Sailkatzaileak neurri multzo ezberdinekin doitzea probatu genuen. Hau da, neurriak lau multzo ezberdinetan banatu genituen, multzo bakoitzean 3.2 atalean azaldutako galdera bakoitzeko eta 3.1.1, 3.1.2 eta 3.1.3 ataletan azaldutako neurri mota ezberdinetarako neurri bat jarri genuelarik. Lehenengo, bigarren eta hirugarren galderako neurrientzat (6) funtzioa erabili dugu eta laugarren galderarako (5) funtzioaren moldapen bat.

- Alde batetik **denbora**-multzoa, 3.1.1 atalean azaldu ditugun denbora luzeena T_l, bataz besteko denbora $\overline{t_l}$ eta erabilitako denboraren desbiderapen estandar $\sigma(t_l)$ neurriekin.
- Bestetik, 3.1.2 atalean, kalkulaturiko $\Delta\alpha_l$ balioa erabiliko dugu **alpha deltak**-multzorako.
- 3.1.2 ataletik ere, β_l neurria erabili da **betak**-multzoa sortzeko.
- Azkenik, 3.1.3 atalean azaldutako d_1 eta d_2 distantziak erabili dira **banaketak**-multzoa sortzeko.

Sailkatzaileen errendimendua neurtzeko, ROC kurba[4] eta ROC kurbaren azpialdeko azalera kontuan hartu dira.

1. irudia. Neurri multzo ezberdinekin ateratako ROC kurba



(a) Banaka (b) Binaka

1a irudiari begiratuz, ROC kurbaren azpialdeko azalerarik handiena lortutako multzoa *banaketak* multzoa izan da, 0.883 balioarekin. Beraren azpitik *betak* (0.737), *denbora* (0.733) eta azkenik *alpha deltak* (0.678) multzoak daude. Dena den, multzo guztiak batuz, marka hobetzen dugula konprobatu dezakegu, *guztiak* multzoa 0.879 markarekin. 1b irudian berriz, sortutako multzoak binaka batu ditugu, multzo berriak bakarka baino hobeagoak direla konprobatzeko.

[4] ROC kurba, erantzun egokien eta alarma faltsuen ehunekoen erlazioa irudikatzen duen erlazioa da.

4. Ondorioak

Lan honek aztertutako neurrien baliagarritasuna aztertzea du helburu, horregatik ez da sailkatzaile on bat doitzera denbora gehiegi eman.

Burututako lanaren arabera, erabilitako denbora neurriak, adostasun neurriak eta erantzunen banaketa neurriak, langileen iruzurrezko portaera antzemateko baliagarriak direla ondorioztatu dezakegu, neurri guztiak batera ROC kurbaren azpialdeko azalera handitzeaz gain, edozein neurri mutzo ezberdin batzerakoan bakarka baino ROC azalera handiagoa lortzen delako.

Honen salbuespen bakarra banaketak gehi betak multzoa da, banaketak bakarrik erabiltzean lortzen den marka berdina lortzen delako.

Neurrien interesa alde batera utzita, oso subjektibodun atazarekin lan egiterakoan crowd bidezko etiketazioak baliogarriak direla konturatu gara, etiketatze lan osteko iruzur detektatze lanak egin behar izan arren. Denbora tarte txiki batean mikro-lan asko bete daitezkelako, adostasun onargarri bat lortuz. Hala ere, deskribatutako neurrietaz lagunduta, detektatutako iruzurrezko langileak egindako lana ezabatzerakoan, etiketen arteko adostasuna igotzen dela konprobatu dugu, eta horren ondorioz, kalitate hobeagoko corpus bat sortu daiteke.

5. Etorkizunerako planteatzen den norabidea

1a eta 1b irudiei begira 4. ataleko ondorioak atera ditugu, baina hau izan daitekenaren susmo bakar bat da. Azken finean lan honetako datuak erabili ditugu bakarrik eta aztertutako neurriak bakarrik kasu honetan horrela erantzutea gerta daiteke. Horregatik, 3.1 atalean azaldutako neurri guztien azterketa sakon bat etorkizunean egiteko asmoa dugu, objektibodun zein subjektibodun etiketazio lan ezberdinekin. Baina subjektibodun lanetan arreta handiagoa jartzea espero dugu, gutxiago aztertu den eremu zail bat delako.

Beste alde batetik, iruzurrezko langileak detektatzea lortu arren, ezin da kalitatezko corpus bat sortu ahal dela ziurtatu. Hori dela eta, ateratako datuekin eredu konputazionalak sortzea ikerketa bide berri bat izango litzateke. Hala eta guztiz ere, corpusa sortzerakoan kontuan izan diren langile kopurua aztertzeko susmoa dugu. Iruzurrezko langileak kentzerakoan datuen arteko adostasuna igotzen den arren, etiketa gutxiago lortzen dira. Horregatik datu kopuraren eta datu kalitatearen arteko oreka lortu behar da.

Azkenik, azaldu ez den arren, gero eta gehien kolaboratu duten langileak gero eta iruzurrezko langileak izateko probabilitatea daukatela dirudi. Gertaera hau etiketatzea ataza neketsua delako izan daiteke, eta etiketa kopuru batetik pasatzerakoan langileek erantzun bera aukeratzerako joera dutela antza ematen du. Hori dela eta, hurrengo saiakeretan langile bakoitzari mikro-lan kopuru maximo bat egiteko ahalmena ezarriko diegu, nekatzeko denboraren probabilitatea murrizteko.

6. Erreferentziak

Artstein, Ron, eta Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34.555–596.

Bennet, E. M., R. Alpert, eta A. C. Goldstein. 1954. Communications through limited response questioning. *Public Opinion Quarterly* 18.303–308.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70.213–220.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20.37–46.

Davies, M., eta J. L. Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics* 38.1047–1051.

Dawid, A. P., eta A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics* 28.20–28.

Dhall, Abhinav, Roland Goecke, Simon Lucey, eta Tom Gedeon. 2011. Acted facial expressions in the wild database.

Eickhoff, Carsten de Vries, Arjen P. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16.121–137.

Eickhoff, Carsten, eta Arjen P. de Vries. 2011. How crowdsourcable is your task? In *Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, Hong Kong, China.

Ekman, Paul. 1992. An argument for basic emotions. *Cognition & emotion* 6.169–200.

Filatova, Elena. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proc. of LREC 2012, Istanbul, Turkey, May 23-25, 2012*, 392–398.

Fleiss, J.L., eta others. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76.378–382.

Gadiraju, Ujwal, Ricardo Kawase, Stefan Dietze, eta Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the ACM CHI 2015, Seoul, Republic of Korea*, 1631–1640.

Gennaro, Rosario, Craig Gentry, eta Bryan Parno. 2010. Non-interactive verifiable computing: Outsourcing computation to untrusted workers. In *Proc. of CRYPTO'10, Santa Barbara, CA, USA*, 465–482.

Gunes, Hatice, eta Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)* 1.68–99.

Gurney, Edmund. 1884. What is an emotion? *Mind* 9.421–426.

Ipeirotis, Panagiotis G., Foster Provost, eta Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proc. of the ACM SIGKDD*, 64–67, New York, USA.

Justo, Raquel, José M. Alcaide, eta M. Inés Torres. 2016. Crowdscience: Crowdsourcing for research and development. In *Proc. of IberSpeech'2016, Portugal*, 403–410.

Klaus, Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Krippendorff, Klaus. 2004. *Content analysis: An introd. to its methodology*. Sage.

——. 2007. Computing Krippendorff's Alpha Reliability. Technical report, University of Pennsylvania, Annenberg School for Communication.

Mohammad, Saif M, eta Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29.436–465.

Ortony, Andrew, eta Terence J Turner. 1990. What's basic about basic emotions? *Psychological review* 97.315.

Raquel Justo, M Inés Torres, José M Alcaide. 2017. Measuring the quality of annotations for a subjective crowdsourcing task. In *Iberian Conference on Pattern Recognition and Image Analysis*, 58–68. Springer.

Rothwell, Spencer, Ahmad Elshenawy, Steele Carter, Daniela iraga, Faraz Romani, Michael Kennewick, eta Bob Kennewick. 2015. Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In *Proc. of Interspeech 2015, Dresden, Germany, September 6-10, 2015*, 2784–2788.

Russell, James A, eta Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality* 11.273–294.

Scherer, Klaus R. 1999. Appraisal theory. *Handbook of cognition and emotion* 637–663.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19.321–325.

Tarasov, Alexey, Sarah Jane Delany, eta Charlie Cullen. 2010. Using crowdsourcing for labelling emotional speech assets.

Valstar, Michel, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, eta Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 3–10. ACM.

7. Eskerrak eta oharrak

Egileok gure esker ona adierazi nahiko genioke Euskal Herriko Unibertsitateari, Espainako gobernuako TIN2017-85854-C4-3-R zenbakidun diru laguntzari eta H2020 Europako Batzordeko SC1-PM15 programako RIA 7 dialdiko 769872 zenbakidun laguntzari, hurrenez hurren, ikerketa hau babesteagatik.