UNIVERSITAT DE BARCELONA

Grau de Lingüística

Treball de Fi de Grau

Curs 2019-2020

GUIA D'ANOTACIÓ DE LA TOXICITAT

Clara Giralt Mirón

Tutor: Mariona Taulé Delor

Barcelona, 12 de Juny de 2020.

Resum:

En aquest treball es fa una proposta d'una guia d'anotació per etiquetar la toxicitat o llenguatge d'odi, amb diferents trets i paràmetres per tal de caracteritzar aquest tipus de llenguatge i poder classificar-lo en diferents graus de toxicitat. A partir d'aquesta proposta s'ha anotat un subconjunt del corpus NewsCom-HS que inclou un total de 1262 comentaris que corresponen a quatre temes: economia, immigració, política i religió. Dels comentaris anotats, 302 són lleugerament tòxics, 144 tòxics i 69 molt tòxics. A més, els resultats de l'anotació del corpus ens permeten observar quins trets són més útils per tal de caracteritzar i classificar el llenguatge d'odi, i també crear un corpus Gold Standard, és a dir, un corpus anotat amb una anotació fiable i de qualitat que s'utilitzarà per tal d'entrenar sistemes de detecció automàtica de la toxicitat.
Paraules clau: llenguatge d'odi, corpus, anotació, Gold Standard.

Abstract:

This work is a proposal for a hate speech annotation guide, with different traits and parameters that allows us to characterize hate speech and help us classificate it in different levels. With this proposal, we have annotated a subset of the corpus NewsCom-HS, which contains a total of 1262 comments that correspond to four different topics: economy, immigration, politics, and religion. Of all the annotated comments, 302 are slightly toxic, 144 are toxic and 69 are very toxic. Also, the results of the corpus annotation using our guide allows us to observe which traits are more useful to characterize and describe hate speech, and to create a Gold Standard corpus, a corpus with a reliable and high quality annotation, which will enable us to train automatic hate speech detection systems.
Keywords: hate speech, corpus, annotation, Gold Standard.

Agraïments

M'agradaria primer agrair a la meva tutora, Mariona Taulé, per la seva confiança i per anar sempre més enllà a l'hora d'ajudar-me durant la realització d'aquest treball. També a ella i a la Toni Martí, per donar-me l'oportunitat de participar en aquest projecte, escoltar-me i tenir en compte les meves propostes.

No tinc suficient espai per donar les mil gràcies que es mereix tot el Servei de Tecnologia Lingüística. M'agradaria agrair especialment a la Montse Nofre, per la seva ajuda amb el corpus i l'estadística del treball, i a en Víctor Bargiela i en Xavier Bonet, per la feinada que han dut a terme amb l'anotació del corpus i la seva ajuda en la creació de la guia d'anotació.

També als meus pares, l'Anna i la Mar, per la seva paciència i comprensió infinita; per ser sempre al meu equip.

A les amigues de sempre, per donar-me la mà quan és fosc.

I a en Xavi, és clar. Per fer-se sentir a prop fins i tot quan és lluny. Gràcies.

Índex

1.Introducció

El discurs d'odi o *hate speech* (també anomenat llenguatge tòxic o abusiu), és el terme més utilitzat per referir-se a qualsevol missatge abusiu o ofensiu, que ataca, denigra o incita a la violència o odi envers a una persona o grup de persones en funció d'unes característiques específiques (com per exemple, la raça, l'ascendència, la religió o la llengua, entre d'altres). Cal tenir en compte també que aquest tipus de contingut d'odi es pot presentar de diferents maneres, ja sigui directament o de manera més subtil, fent ús d'un llenguatge més despectiu o menys ofensiu, etc. Es tracta d'un tema que genera especial interès actualment, entre d'altres motius, perquè la proliferació d'internet, i en gran part l'anonimat que aquest ha pogut comportar, ha fet que el discurs d'odi sigui recurrent a les xarxes socials. Aquest mateix anonimat, i el fet que aquest missatge es pugui viralitzar, aspectes que precisament ho distingeixen de la comunicació *offline*, només fan que accentuar el dany que pot causar, tant a nivell individual com social. A més, el discurs de l'odi sovint s'utilitza contra minories o col·lectius ja de per sí discriminats o oprimits, com poden ser les dones o els immigrants, i per tant l'únic que s'aconsegueix és augmentar o fomentar aquesta discriminació.

S'ha convertit en un problema especialment important per les diferents xarxes socials, que cada cop tenen més pressió per part dels seus usuaris, però també per part d'organitzacions internacionals, perquè abordin i solucionin aquest problema. Diverses xarxes socials, com *Youtube*, *Facebook*, *Twitter* i fins i tot diaris digitals tenen en compte el *hate speech* i presenten normes de comportament per tal d'intentar frenar i eliminar aquest llenguatge tòxic o d'odi[1]. Per exemple, *Youtube* prohibeix el contingut que incita l'odi, que descriu com "contingut que promogui o justifiqui la violència contra persones o grups per motius de raça o origen ètnic, religió, discapacitat, sexe, edat, nacionalitat, condició de veterà de guerra, casta, orientació sexual o identitat de gènere, o que fomenti l'odi per aquests motius"[2]. Però és evident que la gran quantitat de contingut generat pels usuaris fa gairebé impossible la detecció manual del discurs d'odi i, clarament, les normes de comportament i les "llistes negres" de paraules o expressions que sovint es fan servir a les xarxes socials per tal de combatre o evitar aquest llenguatge abusiu són insuficients per a la seva detecció.

Per tant, és important i necessari avançar en la detecció automàtica de *hate speech* i en la creació d'eines que ens ajudin a detectar-lo perquè, tal i com es comenta a Basile et al. (2019: 54): "Aquest tipus de contingut abusiu té un impacte important en el desenvolupament de la societat i podria combatre's amb eines que ens ajudin a detectar-lo automàticament".

L'objectiu principal d'aquest treball és crear una guia d'anotació per anotar el discurs de l'odi o toxicitat en comentaris a notícies. És a dir, elaborar una sèrie de criteris lingüístics que ens ajudin a identificar el llenguatge tòxic i, en concret, el grau de toxicitat. Tot i que el nostre interès principal és el discurs de l'odi també elaborarem una llista de paràmetres o trets que ens ajudin a definir i discriminar els diferents nivells de toxicitat, com per exemple la constructivitat, el sarcasme o l'insult, que mantenen una relació estreta amb la

[1] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
https://www.facebook.com/communitystandards/violence_criminal_behavior
[2] https://www.youtube.com/about/policies/#community-guidelines

toxicitat. Així doncs, l'objectiu principal d'aquest treball és principalment caracteritzar el llenguatge d'odi o tòxic.

Un altre objectiu important del nostre treball és elaborar un corpus anotat amb la toxicitat, el corpus NewsCom-HS. Es tracta d'un corpus creat a partir del corpus NewsCom (Taulé et al. 2019), un corpus utilitzat per anotar el focus de la negació, que està constituït a partir de comentaris *online* a noticies publicades en diaris digitals en castellà, un recurs excel·lent per tal d'estudiar la toxicitat del llenguatge, donat que presenta una comunicació directa entre persones d'origen molt diferent sobre temes d'importància i també sobre temes que poden causar molta controvèrsia. Aquest corpus seria el primer del seu tipus en llengua castellana, perquè la majoria dels corpus anotats amb *hate speech* fins ara han estat majoritàriament en anglès.

Un corpus d'aquestes característiques seria de molta utilitat perquè a més de poder estudiar el discurs de l'odi, ens donaria la possibilitat d'estudiar també com s'expressa l'opinió pública a través dels comentaris a notícies, la connexió que podem trobar entre els comentaris i la connexió entre els articles i els comentaris en resposta. Aquest corpus també seria útil per tal d'avançar en la detecció automàtica de la toxicitat, ja que per entrenar els sistemes de detecció automàtica és necessari tenir un corpus anotat. Per tant, l'anotació del corpus NewsCom-HS seria la meva contribució en una àrea de molt interès actualment, ja que la detecció de la toxicitat podria ajudar a combatre la viralització de missatges que poden incitar a la violència o, fins i tot, a atacar col·lectius o persones en concret.

En aquest treball primer presentarem el marc teòric que ens permetrà situar-nos en l'estudi actual del discurs d'odi i ens proporcionarà uns primers criteris amb els quals començar a treballar (secció 2). A continuació, presentarem la metodologia duta a terme per a l'anotació del corpus (secció 3). En concret, descriurem la guia d'anotació amb els diferents trets i criteris establerts per tal de caracteritzar el discurs d'odi i que ens ajudarà a classificar la toxicitat en diferents graus, així com els resultats de l'anotació del corpus (les proves d'acord entre anotadors) i la creació del *Gold Standard*. Per últim, es presenten les conclusions (secció 4).

En definitiva, amb aquest treball proposem criteris per tal de caracteritzar el discurs de l'odi, crear el primer corpus de toxicitat anotat en llengua castellana i, en general, ajudar i avançar en la detecció automàtica del *hate speech*.

2. Estat de l'art

En aquesta secció comentarem diversos treballs que s'emmarquen en el tema del discurs de l'odi[3], la toxicitat i el llenguatge abusiu. D'aquests treballs, els punts més importants que comentarem seran els problemes que presenten les diferents propostes pel que fa a la definició del discurs de l'odi i els diferents termes que s'utilitzen (secció 2.1); els tipus de corpus que fan servir per analitzar aquest tipus de llenguatge i quines característiques o trets s'utilitzen per tal de caracteritzar-lo (secció 2.2).

2.1. Definicions diferents de discurs d'odi

La detecció del discurs de l'odi és una línia de recerca molt recent i que, per tant, encara presenta diverses problemàtiques. Una d'elles és la falta de consens entre el que es considera discurs d'odi o no i quina terminologia es fa servir per referir-s'hi, especialment en el Processament del Llenguatge Natural (PLN): discurs o missatges d'odi, llenguatge agressiu, llenguatge ofensiu, llenguatge abusiu, entre d'altres.

Podem trobar diferents àmbits des dels quals s'ha estudiat el discurs de l'odi i des del quals s'han proposat diferents definicions, entre els que s'inclouen els estudiats en aquest treball però també en d'altres com la psicologia o la sociologia. Els àmbits que comentarem a continuació són: a) l'àmbit institucional o legal, que inclouria les definicions proposades per diferents organitzacions internacionals (com per exemple les Nacions Unides i el Consell Europeu); b) les xarxes socials (*Facebook*, *Twitter* i *Youtube*) i, per últim, c) l'àmbit més acadèmic o científic i especialment el relacionat amb el PLN. Tal i com es pot veure a les Taules 1, 2 i 3, cada un d'aquests àmbits presenta diferències a l'hora de descriure què és el discurs de l'odi i, fins i tot, podem trobar que no hi ha acord en les descripcions dins d'un mateix àmbit.

Institució	Descripció
Nacions Unides[4]	Discurs d'odi: "qualsevol forma de comunicació oral, escrita o de comportament, que ataca o usa llenguatge despectiu o discriminatori amb referència a una persona o grup basant-se en la seva religió, ètnia, nacionalitat, raça, color de pell, ascendència, gènere o qualsevol altre factor identitari. "
Llei Europea[5]	Discurs d'odi: "Incitació pública a la violència o odi dirigit cap a grups o individus basant-se en certes característiques, entre les que s'inclouen raça, color de pell, religió, ascendència, nacionalitat o origen ètnic."
Consell Europeu[6]	Discurs d'odi: "És un terme usat per descriure un discurs ampli, que és extremadament negatiu i que constitueix un perill per la pau social. Inclou qualsevol expressió que difongui, inciti, promogui o justifiqui el racisme, la xenofòbia, l'anti-semitisme o altres formes d'odi basades en la intolerància".

[3] *Hate speech*, en anglès.

[4] https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

[5] https://ec.europa.eu/commission/presscorner/detail/en/MEMO_18_262

[6] https://www.coe.int/en/web/freedom-expression/hate-speech

Ajuntament de Barcelona/Recomanació nº 15 de la Comissió Europea contra el Racisme i la Intolerància (ECRI) del Consell d'Europa (2015)[7]	Discurs d'odi: "[…] foment, promoció o instigació […] de l'odi, la humiliació o el menyspreu d'una persona o grup de persones, així com l'assetjament, descrèdit, difusió d'estereotips negatius, estigmatització o amenaça pel que fa a aquesta persona o grup de persones i la justificació d'aquestes manifestacions per raons de "raça", color, ascendència, origen nacional o ètnic, edat, discapacitat, llengua, religió o creences, sexe, gènere, identitat de gènere, orientació sexual i altres característiques o condicions personals."

Taula 1: Definicions des de l'àmbit institucional

Xarxa social	Descripció
Facebook[8]	Llenguatge que incita a l'odi: "un atac directe a persones segons el que anomenem característiques protegides: raça, ètnia, nacionalitat, religió, orientació sexual, gènere, identitat de gènere i discapacitat o malalties greus [...] Els comentaris humorístics i socials relacionats amb aquests temes sí que estan permesos".
Twitter[9]	Conducta ofensiva: "No està permès promoure violència o atacar directament o amenaçar altres persones per raons de raça, ètnia, nacionalitat, casta, orientació sexual, gènere, identitat de gènere, religió, edat, discapacitat o malalties greus. "
Youtube[10]	Discurs d'odi: "contingut que promogui o justifiqui la violència contra persones o grups per motius de raça o origen ètnic, religió, discapacitat, sexe, edat, nacionalitat, condició de veterà de guerra, casta, orientació sexual o identitat de gènere, o que fomenti l'odi per aquests motius".

Taula 2: Definicions des de les xarxes socials.

Articles de PLN	Definició
Warner & Hirschberg (2012)	Discurs d'odi: "una forma particular de llenguatge ofensiu que fa servir estereotips per expressar una ideologia d'odi".
Nobata et al. (2016)	Discurs d'odi: "Llenguatge que ataca o denigra a un grup per raons de raça, ètnia, origen, religió, discapacitat, gènere, edat o orientació sexual/identitat de gènere".
Kolhatkar et al. (2018)	Llenguatge tòxic: "[...] inclou llenguatge abusiu, els comentaris ofensius i el discurs d'odi. Un comentari tòxic és un comentari que és probable que ofengui o causi malestar." Dins de llenguatge tòxic s'inclouen també els insults, comentaris sarcàstics que ridiculitzen o que són agressius".
Fortuna & Nunes (2018)	Discurs d'odi: "Llenguatge que ataca o denigra, que incita a la violència o a l'odi contra grups, basant-se en característiques especifiques com l'aparença física, religió, ascendència, nacionalitat o origen ètnic, orientació sexual, identitat de

[7] https://ajuntament.barcelona.cat/bcnvsodi/que-es-el-discurs-d-odi/
[8] https://www.facebook.com/communitystandards/hate_speech
[9] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[10] https://www.youtube.com/about/policies/#community-guidelines

	gènere o altres, i pot produir-se amb diferents estils lingüístics, també de manera subtil o usant humor".
Basile et al. (2019)	Discurs d'odi: "qualsevol comunicació que denigra una persona o un grup de persones basant-se en característiques tals com la raça, color, origen ètnic, gènere, orientació sexual, nacionalitat, religió o altres característiques".

Taula 3: Definicions des de l'àmbit del PLN.

Com es pot observar, es fan servir diferents termes per tal de referir-se al discurs d'odi o *hate speech*, però el més comú, que trobem a definicions dels tres àmbits diferents, és discurs d'odi. A l'àmbit institucional és l'únic terme que trobem, mentre que a les xarxes socials també podem trobar termes similars, com llenguatge que incita l'odi o d'altres com conducta ofensiva. A l'àmbit del PLN el terme discurs d'odi també és el més usat, tot i que amb matisos i més diversitat: Per exemple, Warner & Hirschberg (2012) consideren el discurs d'odi un tipus de llenguatge ofensiu mentre que Davidson et al., (2017) fan distinció entre el discurs d'odi i el llenguatge ofensiu. Waseem et al. (2017) fan servir el terme llenguatge abusiu i proposen una tipologia que distingeix si el llenguatge abusiu va dirigit a un individu o a un grup generalitzat i si el contingut abusiu és explícit o implícit. Kolhatkar et al. (2018) usa el terme toxicitat o llenguatge tòxic, que inclou tant el llenguatge abusiu, els comentaris ofensius i el discurs d'odi, i proposa diferents graus de toxicitat, com veurem més endavant (Secció 2.2).

Més important que els termes usats per referir-se al discurs d'odi són les diferències en les definicions, ja que la definició d'odi que facin servir, els graus d'odi que s'estableixin i què comporti cadascun d'aquests graus condicionarà l'anotació posterior del corpus i en, per exemple, si es tracta d'un missatge sancionable legalment o si és eliminable de la xarxa seguint les polítiques de les diferents companyies, com *Facebook* o *Twitter*. També és important tenir una definició clara per tal de millor l'anotació dels corpus de discurs d'odi que es faran servir per entrenar i avaluar els sistemes de detecció automàtica del discurs d'odi i, per tant, millorar els resultats d'aquests sistemes (Ross et al. 2016).

Seguint la idea de l'estudi sobre detecció automàtica de *hate speech* de Fortuna & Nunes (2018), hem decidit separar les coincidències i diferències entre les definicions de discurs d'odi (Taula 4) i el que aquestes comporten:

Font	**Incitar a l'odi o violència**	**Atacar o denigrar**	**Contra grups específics**	**Humor és acceptable**	**Ús d'estereotips**	**Individualitzat o generalitzat**
Nacions Unides	No	Sí	Sí	No	No	Sí
Llei Europea	Sí	No	Sí	No	No	Sí
Consell Europeu	Sí	No	Sí	No	No	No
Ajuntament de Barcelona/Recomanació Comissió Europea	Sí	No	Sí	No	Sí	Sí

Facebook	Sí	Sí	Sí	Sí	No	No
Twitter	Sí	Sí	Sí	No	No	No
Youtube	Sí	Sí	Sí	No	No	Sí
Warner & Hirschberg (2012)	Sí	No	Sí	No	Sí	No
Nobata et al. (2016)	No	Sí	Sí	No	No	No
Kolhatkar et al. (2018)	No	No	Sí	No	No	No
Fortuna & Nunes (2018)	Sí	Sí	Sí	No	No	No
Basile et al. (2019)	No	Sí	Sí	No	No	Sí

Taula 4: Comparació de definicions.

A l'hora de crear aquesta taula, partim de les quatre característiques proposades per Fortuna & Nunes (2018) per tal de comparar les definicions: 1) El discurs d'odi és incitar a l'odi o violència; 2) el discurs d'odi és atacar o denigrar; 3) el discurs d'odi va sempre dirigit contra grups que comparteixen característiques específiques (religió, ètnia, raça, etc.) i 4) l'humor és acceptable i té un estatus especial. A part d'aquestes característiques proposades per Fortuna & Nunes (2018), n'hem afegit dues de pròpies: 5) El discurs d'odi presenta l'ús d'estereotips i 6) el discurs d'odi pot ser individualitzat o generalitzat, és a dir, pot anar dirigit contra una persona en concret, per exemple, l'autor d'un altre comentari en un fòrum, o contra un grup de persones, normalment un grup minoritzat.

Com es comenta a Fortuna & Nunes (2018), és important destacar que en totes les definicions el llenguatge d'odi va lligat a objectius concrets, és a dir, grups que comparteixen certes característiques, encara que en alguns casos no quedi explícit en la definició. Per exemple, en el cas de Kolhatkar et al. (2018) no queda explícit en la definició, però queda implícit, ja que inclouen el discurs d'odi (que va lligat a grups concrets), entre altres com el llenguatge abusiu o els comentaris ofensius, dins del seu concepte de toxicitat.

També és interessant, a l'hora d'analitzar aquestes dades, fixar-se en les diferències per àmbits. Per exemple, podem veure que les definicions de discurs d'odi de les institucions sovint parlen més aviat "d'incitar a l'odi o a la violència" mentre que les xarxes socials, a l'hora de descriure aquest tipus de llenguatge, parlen més "d'atacar o denigrar". En canvi, a l'àrea del PLN es fan servir les dues i no hi podem trobar una tendència clara.

Hi ha característiques, com la contemplació de l'humor com atenuant, que només es donen en una definició, com és el cas de *Facebook*, una de les xarxes socials. La definició que proposen Fortuna & Nunes (2018), en canvi, també incorporaria l'humor inapropiat com una forma més subtil de discriminació i, per tant, també el marcarien com a discurs d'odi.

L'ús d'estereotips com a definitori del llenguatge d'odi tampoc és massa comú, ja que només es dona en dos casos, un en l'àmbit institucional, a la definició de l'Ajuntament de Barcelona, i l'altre en l'àmbit acadèmic, en concret a Warner & Hirschberg (2012).

També Fortuna & Nunes (2018) els tenen en compte, com una forma subtil de discriminació.

La separació en la recepció d'aquest llenguatge d'odi, és a dir, si va dirigit cap a un grup o un individu en concret, també és una característica present en diferents definicions dels tres àmbits. Waseem et al. (2017) crea una tipologia del llenguatge abusiu a partir de si l'abús va dirigit a un individu concret o un grup generalitzat. Aquesta característica es podria relacionar amb l'abús dirigit contra grups específics, que es troba present en totes les definicions, tant explícitament com implícitament.

Tal i com es comenta a Schmidt & Wiegand (2017), discurs d'odi és un dels termes més usats, tant per ser un terme general utilitzat per referir-se a molt tipus de contingut ofensiu generat a internet com per ser un terme legal en diversos països. També és interessant per l'amplitud que presenta, ja que el discurs pot prendre moltes formes i realitzar-se en qualsevol tipus d'interacció (en xarxes socials, per exemple, pot ser en comentaris, *tweets*, etc) i l'odi pot expressar-se en diferents graus i de diferents maneres.

A l'hora de dur a terme aquest treball hem decidit centrar-nos més aviat en l'àmbit del PNL i, en particular, ens centrarem i basarem els nostres primers paràmetres en el treball dut a terme per Kolhatkar et al. (2018). És per això que hem decidit utilitzar el mateix terme que ells usen, llenguatge tòxic, per sobre d'altres més usats com discurs d'odi o llenguatge abusiu. També ens hem decidit pel terme llenguatge tòxic per ser gairebé sinònim de discurs d'odi però presentar una diferència que ens interessa especialment: el llenguatge tòxic, com podem veure a la definició de Kolhatkar et al. (2018), i al contrari que els termes usats en altres propostes, també abasta comentaris ofensius que no van dirigits cap un grup o individu de característiques concretes o pertanyent a un col·lectiu minoritari.

Per tant, la nostra proposta, més afí al PLN que als altres àmbits, intenta incloure en la seva definició totes les definicions anteriors i optem pel terme toxicitat perquè les englobi. A més, seguint la proposta de Kolhatkar et al. (2018), també ens hem decidit a classificar i anotar la toxicitat en diferents graus (una escala en valors discrets de l'1 al 4) perquè creiem que ens ajudarà a reflectir els diferents matisos de totes les definicions presentades.

2.2. Corpus de HS

En aquesta secció presentem els diferents corpus recopilats que s'han usat en PLN per a l'anàlisi i detecció del discurs d'odi i les característiques que els diferencien (Taula 5). Les característiques que tenim en compte són les següents: la font (és a dir, d'on s'extreuen els textos anotats: *Twitter*, *Facebook*, diaris digitals...), la quantitat de textos anotats, la llengua en què estan escrits els textos i el tipus d'anotació que s'ha dut a terme.

Corpus:					
Article	**Nom**	**Font**	**Quantitat**	**Llengua**	**Anotació**

Warner & Hirschberg (2012)		Yahoo! News (comentaris)	1000 comentaris	Anglès	S'anota el tipus d'odi: anti-semitisme, anti-negre, anti-asiàtic, anti-dones, anti-musulmà, anti-immigrant i altre.
Ross et al. (2016)		Twitter	541 tweets	Alemany	S'anota el "hate speech" (Sí/No), si el tweet s'hauria d'eliminar (Sí/No) i quin grau d'ofensivitat presenta (Escala 1: No ofensiu – 6: Molt ofensiu)
Nobata et al. (2016)		Yahoo! Finances and News (comentaris)	951.736 comentaris	Anglès	S'anota si és abusiu (No abusiu/Abusiu) i, si és abusiu, s'anota quina categoria (odi, llenguatge despectiu o blasfèmia)
Waseem & Hovy (2016)	Corpus collection	Twitter	16.914 tweets	Anglès	S'anota si un tweet és ofensiu o no (Si/No)
Vigna et al. (2017)	Hate Speech Italian Corpus	Facebook (comentaris de pàgines públiques)	17.567 comentaris	Italià	S'anota l'odi en tres categories: Odi fort, odi fluix i no odi i els missatges d'odi s'anoten segons el tipus: Religió, discapacitat, estatus socioeconòmic, polític, raça, sexe, qüestions de gènere i altres.
Davidson et al. (2017)	HateBase Twitter	Twitter	25.000 tweets	Anglès	Anotat per tres o més persones en tres categories: "hate speech", ofensiu però no "hate speech" i ni "hate speech ni ofensiu.

Kolhatkar et al. (2018)	SFU Opinion and Comments Corpus (SOCC)	Articles d'opinió i comentaris d'un diari digital	10 articles d'opinió i 1.043 comentaris	Anglès	Anotació de constructivitat (Constructiu/No constructiu) i toxicitat (Escala 1: No tòxic – 4: Molt tòxic)
Basile et al. (2019)	HatEval dataset	Twitter	19.600 tweets	Anglès	S'anota el "hate speech" (Sí/No), l'objectiu (grup genèric/individu) i l'agressivitat (Sí/No)

Taula 5: Corpus de discurs d'odi.

Com s'observa a la Taula 5, hi ha diferents fonts que s'han fet servir a l'hora de recopilar corpus relacionats amb el discurs d'odi. La font més utilitzada ha sigut *Twitter,* però a part de *tweets* també tenim altres tipus de textos, com comentaris de notícies. També varia la llengua del corpus: la majoria de corpus que existeixen són per a l'anglès, però també en podem trobar per a l'italià (Vigna et al. (2017)) o per a l'alemany (Ross et al. (2016)). La quantitat de textos anotats també és molt variable en els diferents corpus, el corpus més gran és el corpus de Nobata et al. (2016), constituït a partir de 951.736 comentaris, i el corpus més petit, de Ross et al. (2016), està format a partir de 541 *tweets*.

Tots els corpus fan servir diferents anotacions per tal d'anotar el que consideren important en el discurs d'odi. Per exemple, trobem diferents corpus que fan servir una anotació binària (és a dir, que només anoten si presenten discurs d'odi o no), com Ross et al. (2016), Nobata et al. (2016), Waseem & Hovy (2016) i Basile et al. (2019). Altres, per tal d'anotar el discurs d'odi, fan servir una gradació: Vigna et al. (2017) i Davidson et al. (2017) fan servir un sistema de tres graus, Kolhatkar et al. (2018) fa servir un sistema de quatre graus i Ross et al. (2016) usa un sistema de sis graus. En altres corpus s'utilitza un sistema per categories. Per exemple, a Nobata et al. (2016) usen tres categories diferents: odi, llenguatge despectiu i blasfèmia; mentre que a Warner & Hirschberg (2012) i Vigna et al. (2017) usen sistemes de set i vuit categories respectivament, depenent del grup específic a qui va dirigit l'odi. També hi ha corpus on s'anoten altres característiques que poden estar relacionades amb el discurs d'odi, com poden ser l'agressivitat (Basile et al. (2019)) i la constructivitat (Kolhatkar et al. (2018)).

3. Metodologia

3.1. Corpus NewsCom-HS.

Un dels objectius d'aquest treball és l'anotació d'un corpus amb informació sobre la toxicitat, el corpus NewsCom-HS, que és el primer del seu tipus en llengua castellana, doncs la majoria de corpus anotats amb toxicitat són en llengua anglesa. Aquest corpus està creat a partir del corpus NewsCom, un corpus utilitzat per anotar la negació i el seu focus (Taulé et al. 2019).

El corpus NewsCom consisteix en 2955 comentaris, publicats en resposta a nou articles de diferents diaris digitals espanyols des d'agost del 2017 a maig del 2019. Aquests articles tracten nou temes diferents: immigració, política, tecnologia, terrorisme, economia, societat, religió, refugiats i immobiliària. Els comentaris es van seleccionar en l'ordre en què apareixen cronològicament i es van eliminar els duplicats, deixant només els originals i únics. Els comentaris estan escrits en llenguatge informal i, per tant, és molt possible trobar comentaris que no siguin gramaticals o continguin erros sintàctics.

Considerem que els comentaris a notícies online són un molt bon recurs a l'hora d'estudiar la toxicitat, ja que és una comunicació directa sobre temes candents i que poden ser molt controvertits entre persones d'origen molt diferent. Els primers temes anotats amb toxicitat del corpus NewsCom-HS en són quatre: economia, immigració, política i religió. Aquests temes, precisament, van ser seleccionats tenint en compte la potencial visceralitat dels seus comentaris que, sumat a l'anonimat que permeten aquestes seccions de comentaris, és possible que ens aportin respostes molt interessants a l'hora d'estudiar la toxicitat.

Un corpus d'aquestes característiques no només és útil per estudiar la toxicitat, sinó que també permetrà analitzar com s'estudia l'opinió pública a través de comentaris, la connexió entre la notícia publicada i els seus comentaris i la connexió entre els comentaris mateixos (però que no es tractaran en aquest treball). Aquest corpus també es podrà utilitzar per entrenar i avaluar sistemes de detecció automàtica de toxicitat.

3.2. Definició de la tasca

La tasca que durem a terme consisteix en l'anotació de la toxicitat en cada comentari dels quatre temes seleccionats del corpus NewsCom-HS. Per tal de dur-la a terme cal primer definir la metodologia d'anotació, és a dir, establir l'etiquetari que s'utilitzarà, els criteris per aplicar-lo i el procés d'anotació (la manera en què es farà).

En aquest treball es realitzaran dues tasques que es corresponen amb els objectius principals del treball:

1) Definició de la metodologia d'anotació i, en concret, la guia d'anotació;
2) Anotació del corpus NewsCom-HS amb toxicitat seguin els criteris establerts a la guia.

3.3. Guia d'anotació

1) Proposta inicial de Kolhatkar et al. (2018)

Per abordar la tasca d'anotació del corpus amb toxicitat i veure a què ens enfrontàvem vam fer una primera anàlisi dels exemples de dos dels temes seleccionats, el d'economia i el d'immigració, aplicant la proposta d'anotació de Kolhatkar et al. (2018). Aquests autors anoten la toxicitat en quatre graus diferents (1-4) i també la constructivitat amb un valor binari (SÍ/NO). Hem decidit optar per aquesta proposta perquè els criteris i paràmetres que proposen ja han sigut utilitzats abans i han donat bons resultats: el percentatge d'acord en l'anotació de la constructivitat en una mostra aleatòria de 100 anotacions va ser del 87.88%, mentre que el percentatge d'acord en l'anotació de la toxicitat en una mostra aleatòria de 100 anotacions va ser del 81.82%. Creiem també que l'anotació del corpus amb el tret de constructivitat ens pot ajudar a l'hora d'anotar els graus de toxicitat.

Una altra raó per seguir aquesta proposta ha sigut la manera com han obtingut la seva definició de constructivitat. En comptes d'elaborar una definició pròpia o fer servir una de ja existent, van decidir crear una enquesta en línia preguntant a 100 usuaris què creien que era un comentari constructiu i, a partir de les respostes, van aconseguir les característiques principals per tal de definir la constructivitat. És a dir, es tracta d'una definició de constructivitat creada amb *crowdsourcing*[11].

Les característiques que van recollir per distingir entre comentaris constructius i no constructius són les següents:

Comentaris constructius	**Comentaris no constructius**
Creen diàleg	No tenen gaire contingut
Són comentaris rellevants	Assignen culpa
Presenten evidència	Presenten reaccions emocionals
Presenten solucions	Són sarcàstics
Presenten noves perspectives	No presenten evidència
Presenten experiència pròpia	Són degradants
	Són excessivament afalagadors

Taula 6: Característiques dels comentaris constructius i no constructius segons Kolhatkar et al. (2018)

La definició de constructivitat que proposen és la següent i és la que adoptarem en la guia d'anotació:

"Els comentaris constructius pretenen crear diàleg civilitzat a partir d'observacions que són rellevants a l'article i no pretenen només provocar una resposta emocional; normalment es refereixen a punts específics de l'article i aporten evidència apropiada" (Kolhatkar et al. (2018: p. 12)).

Pel que fa a l'anotació de la constructivitat en els 1121 comentaris també es va dur a terme amb *crowdsourcing*[12], i el percentatge d'acord obtingut en una mostra de 100 comentaris

[11] Van utilitzar la plataforma SurveyMonkey: https://www.surveymonkey.com

[12] Per a l'anotació dels comentaris, tant amb constructivitat com amb toxicitat, va fer servir la interfície *CrowdFlower* (https://www.crwdflower.com).

va ser del 87.88%. Hagués estat interessant, però, saber el percentatge d'acord sobre el total de les anotacions i tampoc queda clar què en fan amb els casos de desacord.

Pel que fa a l'anotació dels graus de toxicitat, també van seguir un procés d'anotació *crowdsourcing*, van utilitzar la mateixa interfície *CrowdFlower* per anotar els diferents 1121 comentaris demanant que els usuaris els classifiquessin en els quatre graus de toxicitat possibles.

Les característiques dels quatre graus de toxicitat que Kolhatkar et al. (2018) proposen a l'hora d'anotar són les següents:

Molt tòxic (4)	**Tòxic (3)**	**Lleugerament tòxic (2)**	**No tòxic (1)**
Presenten llenguatge ofensiu o abusiu	Sarcàstics, crítiques no constructives	Podrien ser considerats tòxics	Comentaris que no presenten toxicitat.
Presenten atacs personals i/o insults	Presenten burla o ridiculitzen	Expressen frustració o ràbia	
Són despectius o degradants	Discrepen agressivament		
	Presenten bromes inapropiades		

Taula 7: Característiques dels quatre graus de toxicitat segons Kolhatkar et al. (2018)

2) Reformulació de la guia d'anotació

Al dur a terme aquesta primera aproximació de l'anotació del corpus NewsCom, ens vam adonar que aquests paràmetres establerts per Kolhatkar et al. (2018) presenten diversos problemes. El més important és que els criteris són massa concisos, que cal definir més acuradament cadascun d'aquests graus de toxicitat i, en especial, els diferents paràmetres o trets que els defineixen: què s'entén per sarcasme, com el diferenciem de la ironia, com establim quan és un missatge més o menys ofensiu. Es tracta de trets tots ells molt subjectius que depenen massa de la interpretació de cada anotador.

Per exemple, el primer paràmetre del grau 2 de toxicitat és que els comentaris poden ser considerats tòxics per alguna gent o en alguns contextos, que és un criteri molt subjectiu i que podria abastar molts comentaris diferents. També trobem diversos problemes a l'hora de diferenciar el sarcasme, que correspon al grau 3 de toxicitat, de la ironia o l'humor, que són característiques que no tenen grau assignat. Igual que el sarcasme, l'agressivitat també és un concepte molt subjectiu i que pot presentar diversos graus, fent així molt difícil la seva classificació en aquesta escala proposada per Kolhatkar et al. (2018). De fet, són trets que també presenten gradació i, per tant, difícils de definir.

Un cop observats aquests problemes i a partir de l'anàlisi corresponent, hem considerat necessari dur a terme una reformulació de la guia d'anotació inicial. Aquesta reformulació inclou una ampliació dels criteris d'anotació i la incorporació de més trets a l'anotació que ens ajudin a distingir i classificar millor els diferents graus de toxicitat. Els nous criteris de definició de graus i els nous trets a anotar s'han creat a partir dels problemes observats en la primera anotació del corpus, que s'han discutit en reunions dutes a termes

amb els anotadors on s'han acordat criteris més detallats i on s'han intentat resoldre els casos més problemàtics.

La reformulació d'aquests criteris i la inclusió de nous trets en el procés d'anotació es presenten a la secció següent (3.3.1).

3.3.1. Etiquetari i criteris d'anotació

En aquesta secció comentarem els nous criteris i trets d'anotació que proposem per tal de millorar la classificació dels comentaris del corpus en els diferents graus de toxicitat. Pel que fa a la constructivitat, seguim basant-nos en la definició i criteris establerts en la primera versió de la guia d'anotació. Aquesta nova versió de la guia d'anotació es crea com a resposta als problemes i dubtes que han sorgit als anotadors durant l'anotació inicial del corpus i que, per tant, implica que ha sigut una tasca que s'ha anat ajustant a mesura que avançàvem en l'anotació del corpus.

Els nous trets que hem afegit a l'anotació en són cinc: sarcasme, burla o ridiculització, insults, argumentació i llenguatge negatiu o tòxic. Tots ells són trets binaris amb el valor SÍ/ NO, és a dir, només es marcarà si el presenten o no. A la guia, cada un d'aquests nous trets van acompanyats de diferents criteris i característiques per tal de poder descriure'ls amb suficient claredat, saber com i quan poder-los aplicar en l'anotació, i per tant que ens ajudin en la classificació de toxicitat.

A continuació es presenta primer com es defineixen cadascun d'aquests tres i després, tenint en compte aquests trets, quins són els criteris per anotar els graus de toxicitat.

1. Sarcasme:

El sarcasme és un tret que pot ser difícil d'anotar, especialment per la seva subjectivitat i pels problemes que pot presentar a l'hora de distingir-lo de la ironia o de l'humor. Per tal d'ajudar en la seva detecció hem proposat una llista de recursos expressius que ens poden ajudar a classificar què considerem sarcasme i què no ho és (Taula 8).

Recursos expressius:	**És sarcasme**	**No és sarcasme**
Ironia	Si és ofensiva o degradant (1)	Si és suau, no presenta atacs (*Ironia o sarcasme blanc*) (6)
Preguntes retòriques	Si van dirigides a un altre usuari o utilitzen un to burleta (2)	Si no tenen destinatari concret (7)
Metàfores	Si són sarcàstiques o crítiques contra algú (3)	Si fan servir una ironia suau o no presenten atac ni ràbia (8)
Frases fetes o expressions comunes	Si van acompanyades de crítica o ràbia (4)	Si no van acompanyades de crítica, tot i que poden tenir una certa càrrega emotiva o expressiva (9)
Bromes	Si poden resultar ofensives per algun col·lectiu (5)	Si només tenen intenció de fer riure (10)

Taula 8: Sarcasme

Exemples de comentaris sarcàstics:

(1) [Política #141]: *Agradecer a ERC que nos haya librado de un presidente del Senado IMPRESENTABLE no tiene precio... Esperemos que Manuel Cruz sepa aplicar alguna sensatez filosófica en una Cámara Altamente Inútil... Y todos contentos... La nave va, la nave va...*

(2) [Política #69]: *un golpe de estado fallido fue lo que realizó Tejero, ¿sabe ver las diferencias?*

(3) [Economia #120]: *Ya era hora!! Era pagar a la gasolinera por ir en bicicleta. Ahora deberían mejorar el sistema tarifario y las puertas giratorias modelo Trini... Imagino q eso no toca.... para abrirles la cartera ya está la clase media! ¿¿¿¿*

(4) [Política #104]: *El puto gallego haciendo de las suyas. El mierda de Kent metièndosela sin vaselina a sus votantes. Nada nuevo cara al sol.*

(5) [Política #210]: *son el pueblo elegido*

Exemples de comentaris no sarcàstics:

(6) [Economia #20]: *que pregunten también en Chile por los planes privados de pensiones, creo que están muy contentos*

(7) [Immigració #275]: *¿Se puede transferir dinero del contribuyente europeo al tiempo que ellos puedan tener presupuestos en los que sigan comprando armamentos? ¿Se puede invertir dinero en educación al tiempo que ellos puedan decidir que las niñas quedan excluidas a cierta edad juvenil? ¿Se puede educara las mujeres para que tengan el control sobre la maternidad, al tiempo que ellos sigan imponiendo sus postulados religioso-culturales que las siga sometiendo? Obviamente no. Y como es muy dudoso que estén dispuestos a ceder una parte de la soberanía para que Occidente les arregle los problemas, todo esto de la ayuda, como hasta ahora, no serán más que parches*.

(8) [Immigració #163]: *Vaya joya de vecina, te compadezco. Espero que su marido no lleve bigote, esos son los peores. Un saludo*

(9) [Economia #107]: *Todo con tal de estrujar a la gallina de los huevos de oro, que es el trabajador.*

(10) [Religió #73]: *a ver, de qué vas? Con goku no se mete nadie, nadie.*

És important destacar que l'ús de malnoms, noms distorsionats o apel·latius pejoratius es pot considerar sarcàstic i s'ha de marcar com a tal. Per exemple:

(11) [Política #]: *Tranquilo Mariano, a PPedro no le importa.*

(12) [Política #]: *Está claro que TV3% (una tele pública) hizo propaganda del referendum cuando estaba prohibido*.

Però també és important tenir en compte, a l'hora de classificar el sarcasme en graus de toxicitat, que hi ha ironia o sarcasme que no pretén ofendre o causar molèsties, sinó que

es fa servir, més aviat, com a recurs estilístic. L'anomenarem *ironia o sarcasme blanc* (Veure Taula 8) i comportarà un grau menor de toxicitat. Trobem exemples a:

(13) [Política #158]: *eso sí que es un mantra.*

(14) [Immobiliària #40]: *Ah, ¿pero es que había bajado del 25?*

2. Burla o ridiculització:

La burla o ridiculització pot ser un tret difícil de classificar, també per la seva subjectivitat i perquè implica gradació. Per tal de fer-ne una millor classificació, presentem una llista de recursos expressius que ens ajudaran a decidir què és burla o ridiculització i què no ho és.

Recursos expressius:	**És burla**	**No és burla**
Crítiques	Contra un grup o col·lectiu determinat: grup ètnic, racial, nacional, identitari, religiós, ideològic, etc. (15)	A la gestió d'un partit polític (21)
		Generalitzades a tots els usuaris i/o sense un destinatari concret (22)
Comentaris	Humiliants o deshumanitzadors (16)	
Atacs	Contra un grup o col·lectiu determinat: grup ètnic, racial, nacional, identitari, religiós, ideològic, etc. (Veure Altres factors) (17)	
	A la intel·ligència i/o les habilitats lingüístiques de persones concretes (18)	
	Aparentment neutres (o amb un grau baix d'agressivitat) però que es fan de manera reiterativa (19)	
Discrepàncies		Amb altres usuaris sense insults i/o agressivitat explícita (23) (24)
Insults o desqualificacions	Explícits i ofensius contra altres usuaris, normalment acompanyats de burla i/o atac (20)	No humiliants, tot i tenir una connotació negativa (25)

Taula 9: Burla o ridiculització

Exemples de comentaris amb burla o ridiculització:

(15) [Política #34]: *y por que no pueden utilizarse? No me lo digas: la sagrada unidad de España es más importante que toneladas de mierda y corrupción*

(16) [Política #86]: *Lo que los catalanes queremos es menos porras y menos orangutanes infectados de evola pegando a poblacion indefensa. Estos gestos que se los metan por donde quieran.*

(17) [Religió #17]: *Boicot a los cristofrikis.*

[Immigració #162]: *Lógico, un arabe con millones por las orejas no le cuesta un duro al estado. Un marroqui que no tiene donde caerse muerto ya cobra, subvenciones y es un gasto sanitario. Si no eres capaz de ver la difererencia...*

(18) [Economia #144]: *Si usted se molestase en escribir mejor, seria mas fácil que comprendiésemos sus comentarios. "njo" "nno" "lred"*

(19) [Religió #180]: *Te he humillado con argumentos y tú te sales por la tangente una y otra vez con tal de no reconocer tu error... Para mí y para cualquiera está claro quien lleva razón y quien es un niñato trol que en cuanto le cazan en una parida (que es lo que has escrito) intenta escabullirse con chistecillos y bromitas. Creerás que engañas a alguien... Y en realidad el único que se engaña eres tú.*

(20) [Política #104]: *El puto gallego haciendo de las suyas. El mierda de Kent metièndosela sin vaselina a sus votantes. Nada nuevo cara al sol.*

Exemples de comentaris sense burla o ridiculització:

(21) [Política #113]: *El pp cumpliendo algo de lo que dice ? El psoe siendo socialista ? pero de que hablan.*

(22) [Política #70]: *Entre mentirosos anda el juego*

(23) [Política #13]: *¿Pasa algo con la Complu? ¿Te parece gracioso? De qué vas.*

(24) [Economia #178]: *Si quieres te hago presupuesto, parece que no puedes encontrar buenos precios.*

(25) [Política #112]: *Va a meterse P.Sánchez en una reforma de la constitución viendo como salió la reforma de un simple estatuto de autonomía. Y para darle más poderes a los catalanes, claro que sí. Os dejáis engañar...*

Cal tenir en compte que per expressar burla no és necessari fer servir un llenguatge groller o paraules malsonants, com podem veure en el següent exemple:

(26) [Economia #176]: *¿3kw 18.000 Eur? Ahhhhh. ¿Baterías cambiadas totalmente en 10 años? Ahhhhh. Oiga, yo me dedico a esto ¿Y vd.?*

En cas que considerem que el comentari presenta burla o ridiculització, haurem de marcar un grau de toxicitat 2 o 3, depenent de la intensitat de la burla o ridiculització que presenti. Per exemple:

(27) GT2: [Política #136]: *Jajaja... No hay como mirarse el ombligo. Sois los mejores esta faceta, junto con el victimismo*

(28) GT3: [Immigració #103]: *facha fascismo nazis blablabla*

3. Insults

Els insults són, en general, fàcils de detectar i, per tant, ens serà més fàcil classificar aquest tret. Per exemple, podem trobar insults en aquests comentaris del corpus:

(29) [Religió #232]: *Como ateo cada vez que me topo con algun religioso, de estos fervientes que intentan hacer proselitismo o hablan de su fe y sus mierdas litúrgicas dando por sentado que me importan un carajo... Pues en esos casos sonrío y asiento educadamente con la cabeza. Pero pienso muy fuerte: "Pero que **subnormal** eres, **hijo de puta"***

(30) [Política #70]: *¿Y no se será que más que que ellos sean unos genios, la base que les vota son directamente **subnormales**?*

Però un punt important a tenir en compte és que els insults els hem de considerar des del punt de vista de l'insultador, doncs una paraula que, canònicament, no seria un insult es pot utilitzar com a tal i, per tant, hauríem de marcar el tret d'insult amb el valor SÍ (i marcar un grau de toxicitat 3). Per exemple:

(31) [Immigració #95]: *No es racismo, **progre**, es sentido común*

En aquest cas, trobem "progre", de progressista, que no seria considerat un insult per a moltes persones, però que aquí s'utilitza amb clara intenció d'ofendre.

Quan anotem el comentari amb el tret 'Insult=SÍ' aleshores el comentari serà tòxic i el grau de toxicitat oscil·larà entre 3 i 4 en funció del tipus d'insult, si és molt ofensiu, si n'hi ha més d'un, aleshores l'anotarem amb un grau de toxicitat 4. L'exemple (31) l'anotaríem amb grau de toxicitat 3 i l'exemple (32) amb toxicitat 4:

(32) [Política #62]: *¿Y no se será que más que que ellos sean unos genios, la base que les vota son directamente **subnormales**?*

4. Llenguatge negatiu o tòxic

Per llenguatge negatiu o tòxic ens referim a l'ús d'insults dirigits cap a una persona, paraulotes o paraules malsonants, comentaris fets amb mala fe o comentaris amb connotacions negatives. Per tal de saber si un comentari presenta llenguatge tòxic o no, una possible estratègia és pensar si el comentari podria aparèixer al faldó de televisió en directe: si creiem que el comentari no podria aparèixer al faldó d'un programa televisiu sense retocar-lo (per exemple, eliminar paraules malsonants), l'anotem com a llenguatge tòxic (exemples (33), (34) i (35)).

(33)[Religió #250]: *es un estudio hecho en EEUU y los estereotipos son los correspondientes a ese país, donde sí hay una percepción negativa de los ateos hasta el punto de (según encuestas) ser la característica menos deseable en un candidato a*

*cargo público. Por estos lares diría que a la gente en general se la bufan bastante las creencias de terceras personas en tanto que no den el **coñazo** con ellas.*

(34)[Economia #24]: *Si, una **puta mierda** el sistema sanitario americano, pero eso no invalida lo que dije.*

(35)[Política #39]: *Nos **jodemos** todos. Rajoy le pega las tortas a Sánchez en la cara de todos los españoles.*

En els casos en què s'anota el comentari amb el tret de 'llenguatge tòxic/negatiu=SÍ' el comentari també s'anotarà com a tòxic i el grau serà 2, 3 o 4 en funció de la intensitat del contingut, dels insults, paraules malsonants, sarcasme, etc.

5. Argumentació

L'argumentació també és un tret fàcil de detectar. Considerarem que un comentari presenta argumentació quan detectem que tingui la intenció d'autoexplicar-se, que fonamenti la seva opinió amb dades i/o que creï diàleg (exemples (36), (37) i (38).

(36) [Economia #26]: *Para incentivar las inversiones hace falta seguridad jurídica y no que se cambie el "marco regulatorio" cada 3 meses. El PP es ecologista en la oposicion y en el poder se carga todo el sector*

(37) [Economia #69]: *Creo que en Chile o Argentina tienen un sistema obligatorio privado de pensiones con aportaciones de los trabajadores y al llegar la hora de cobrarlo están cobrando menos de lo que les habían prometido, siendo incluso inferiores al sistema público, porque las empresas gestoras se quedan con una parte, las empresas gestoras usan las aportaciones para enriquecerse ellas y no los futuros jubilados.*

(38) [Política #21]: *Lee mejor o lee desinteresadamente. Ni nihilismo (he dicho actuar políticamente al margen del voto) Ni todos iguales (he dicho IU fuerte, por ejemplo).*

Aquest tret està molt relacionat amb el de constructivitat, tan és així que si anotem el comentari amb el tret 'Argumentació=SÍ' implica que el comentari és constructiu i, per tant, l'hem d'anotar com a tal. A més, també ens ajuda a delimitar el grau de toxicitat, perquè un comentari constructiu mai no s'anotarà amb un grau de toxicitat 4.

6. Intolerància:

Anotem amb el tret d'intolerància els comentaris que ataquen o denigren a algú basant-se en certes característiques com el seu gènere, orientació sexual, ètnia, nacionalitat o religió, seguint els paràmetres que ofereix l'Oficina per la No-Discriminació de l'Ajuntament de Barcelona[13], que fa distinció entre discurs intolerant (o no sancionable) i discurs sancionable. En base a la Piràmide de l'odi dissenyada per l'Oficina per la No-Discriminació de l'Ajuntament de Barcelona, podem adaptar el model a la nostra escala

[13] https://ajuntament.barcelona.cat/bcnvsodi/que-es-el-discurs-d-odi/

de toxicitat quant als missatges d'odis percebuts als comentaris de les notícies anotats (Taula 10).

No sancionable	**Sancionable**
Estereotips i prejudicis lligats a idees supremacistes (42)	Violència simbòlica o incitació a la violència; insults directes per pertinença a certs col·lectius (47)
Deshumanització, altredat o cosificació (43)	
Idees homonacionalistes[14] o *pinkwashing*[15] (44)	
Estereotips, rumors, bocs expiatoris, assetjament i amenaces (45)	
Ridiculització per motius de pertinença a certs col·lectius (46)	

Taula 10: Intolerància

(42) [Immigració #94]: *Por la simple razón de que nosotros tenemos los medios, el dinero, la cultura y la educación para evitarlo y los gobiernos africanos, no*

(43) [Immigració #104]: *Ahora, que me den una buena razón por la que es bueno acoger a miles de personas de otra cultura, otra educación, otra etnia... para tener que darles de comer.*

(44) [Immigració #108]: *Estamos en nuestro derecho de no desear que en España entren personas que por ejemplo tienen en su cultura la homofobia totalmente incrustada, cuando nosotros es algo que estamos superando. Y como con eso, muchas cosas.*

(45) [Immigració #84]: *En Euskadi muchos de ellos cobran 1000, 2000, o 4000 € de RGI recurriendo a toda clase de subterfugios. El escándalo ha llegado a ser tal que ni los políticos ni los medios, comprados por estos, han podido taparlo y la gente está muy caliente. Claro, el RGI también lo pueden cobrar un jubilado... la diferencia es que el jubilado ha trabajado toda su puñetera vida y le van a dar 50 o 100 € de complemento*

(46) [Religió #249]: *Me importa un pimiento lo que piensen de mi unos tios que hablan con un amigo imaginario.*

[14] Terme originalment proposat per la investigadora en estudis de gènere Jasbir K. Puar per referir-se als "processos pels quals certs poders s'alineen amb les reivindicacions del col·lectiu LGBTI amb la finalitat de justificar posicions racistes i xenòfobes, especialment en contra de l'Islam, recolzant-les sobre els prejudicis que les persones migrants han de ser forçosament homòfobes i que la societat occidental és completament igualitària. D'aquesta forma, es fa ús de la diversitat sexual i els drets LGBT per sostenir postures en contra de la immigració, sent cada vegada més comuna entre partits d'ultradreta". (Viquipèdia: https://ca.wikipedia.org/wiki/Homonacionalisme)

[15] Pinkwashing (rentat rosa o rentat d'imatge rosa): "Varietat d'estratègies polítiques i de màrqueting dirigides a la promoció de productes, empreses o institucions, apel·lant a la seva amabilitat cap al col·lectiu LGBT". S'utilitza especialment per referir-se a les polítiques de l'estat d'Israel, però també s'utilitza per polítiques d'altres països i en altres àmbits no polítics. (Viquipèdia: https://ca.wikipedia.org/wiki/Pinkwashing)

(47) [Immigració #272]: *Joder es terminar las elecciones y que no paren de venir. Mandenlos de vuelta inmediatamente que awui en Cádiz estamos hartos que nos están conquistando. Donde está el ejército para defender nuestras fronteras?*

7. Agressivitat:

L'agressivitat es caracteritza per "l'ús de la violència o el desig d'exercir-la de manera conscient o inconscient[16]", és un factor que implica una alta toxicitat, encara que no presentin altres trets com la burla o el sarcasme, i que per tant hem de tenir en compte a l'hora d'anotar. Trobem agressivitat en els exemples (48) i (49):

(48) [Economia #66]: *Pues si, al final la única opción va a ser cortar otra vez cabezas. Pero hay que tener cuidado, esta vez no les vamos a pillar tan desprevenidos, han aprendido de las revoluciones francesa y rusa. Por algo en los USA existe un ejército interior que se llama la Guardia Nacional.*

(49) [Immigració #293]: *Aquí somos suficientes y muchos más que nunca antes en la historia. Si es que necesitásemos trabajadores cualificados (hay 14-15% de paro en España…). Los que vengan que lo sean, y entren legalmente y dispuestos a integrarse mientras les dura su permiso de trabajo. Sobre en son de paz, no directamente a destruirnos y volarnos por los aires, violarnos, robarnos etc...Ahora los drones, los misiles y demás pueden ayudar a defendernos con amplitud*

8. Altres trets:

Hi ha altres trets que, tot i no ser part dels trets binaris que anotarem, ens poden ajudar a l'hora d'anotar la constructivitat i la toxicitat.

a) Pel que fa a la constructivitat, podem tenir en compte la llargada del comentari i si conté preguntes, ja que tot i que un comentari curt pot ser constructiu (39), si és llarg (40) o si conté preguntes (41) és més possible que ho sigui.

(39) [Economia #2]: *Inversión cara, pero muy rentable, a la que se le puede sacar el máximo ahorro en combinación con fotovoltaica.*

(40) [Economia #6]: *Es cierto que donde estoy el invierno es suave y se que tendré un déficit de generación en los meses de menos insolación (diciembre y enero). En los meses de verano aumenta el consumo también por el aire acondicionado pero hay mayor insolación por lo que el déficit es bastante menor. No se puede dimensionar para cubrir todos los picos porque se dispara el precio. Aunque también te digo que si el consumo de 4000kwh se te dispara a 16000kwh por usar bomba de calor, algo mal pasa. O bien el aparato es de muy poca eficiencia (mírate un inverter), o bien estás en un clima superextremo (muchísimo frío en invierno y muchísimo calor en verano), o bien os pasáis de la cuenta al fijar las temperaturas de referencia (demasiado frío en verano y demasiado calor en invierno). También puede ser un poco de todo.*

[16] Diccionario de Filosofía (en castellà). 1a. Barcelona: SPES Editorial (edició especial per a RBA Editoriales), 2003, p. 3 (Biblioteca de Consulta Larousse).

(41) [Política #37]: *¿cómo sabe la cifra exacta cuando no se permite un referéndum legal y vinculante? ¿Da miedo el permitir una votación cuando sabe que somos minoría?*

b) Per tal d'ajudar-nos a classificar la toxicitat, podem tenir en compte diferents trets, com poden ser la supèrbia, el paratext (o context), els disfemismes o l'ús de majúscules, que es descriuen a continuació.

Supèrbia:

La supèrbia, que definim com "Excessiva estima de si mateix amb menyspreu dels altres[17]" també és un factor que ens podria indicar cert grau (lleu) de toxicitat, correspondria al grau de toxicitat GT2 (exemples (50) i (51)).

(50) [Economia #15]: *Seguro que te experiencia es abrumadoramente superior a la mia, desde luego...*

(51) [Religió #145]: *Los ateos somos conscientes de las deficiencias intelectuales de los creyentes y procuramos tratarles como si no fueran imbéciles... Por aquello del llevarse bien.*

Paratext:

El context i el paratext són factors a tenir en compte a l'hora d'assignar el grau de toxicitat. Cal tenir en compte el text de la notícia i les imatges que l'acompanyen, ja que molts comentaris hi fan referència. Per exemple:

(52) [Religió #10]: *La misma **imagen** con que ilustran el artículo lo dice todo: los dos ladrones al lado del "Maestro": uno creyente y otro incrédulo, ¿Cuál fue más irrespetuoso?... tranquilos...*

(53) [Política #74] *En **esa imagen** se lo deja claro...que sí,que sí,que por un oído me entra y por el otro me sale.*

A l'hora d'anotar, però, no tindrem en compte el context dels comentaris que es contesten els uns als altres i els analitzarem per separat. Per exemple, en el següent fil de comentaris:

(54) [Política #18]: *Como no les ha gustado a los indepes Iceta y probablemente tampoco estos dos, yo propondría a Tomás de Torquemada, un hombre conocido por su tolerancia. Seguramente será del agrado de los indepes.*

(55) [Política #19]: *ese no está en el TS?*

(56) [Política #20]: *No, está en Waterloo*

Si tinguéssim en compte el context dels comentaris, el més probable és que marquéssim els comentaris (55) i (56) com tòxics, al presentar burla, però al no tenir en compte aquesta

[17] *Diccionari de la llengua catalana de l'IEC*. Barcelona : Institut d'Estudis Catalans. https://dlc.iec.cat/

relació entre ells és molt difícil veure la burla, així que seran classificats com a grau de toxicitat GT1.

Disfemismes:

Tenim en compte els disfemismes (ús lingüístic que consisteix a al·ludir objectes, persones, fets, etc, mitjançant formes grolleres o despectives[18]) i els marquem amb cert grau de toxicitat perquè ens presenten una versió distorsionada de la realitat. Per exemple a:

(57) [Religió #157]: *hombre, aquí otro sin tele y q no* ***ha mutilado a la niña*** *al nacer.... Al final no somos TAN pocos como creemos.* (fa referència a les arrecades)

(58) [Immigració #4]: *Rescatado no, sorprendidos en las costas de Cadiz. A ver si llamamos a las cosas por su nombre. Y para llamarlo correctamente diré, ya empieza el cachondeo y la* ***invasión*** *de todos los veranos, expulsiones 30 al mes, llegadas 500 al día.*

Ús de majúscules:

L'ús de majúscules no sempre implica toxicitat, però sí que en alguns comentaris es fan servir per expressar ràbia, frustració o per simular que es crida, casos en els que sí implicaria cert grau de toxicitat (exemples (59) i (60)):

(59) [Política #87]: ***FALSO; ABSOLUTAMENTE falso****, hace falta mayoria del 90%, y ademas, podemos ha dicho que cualquier reforma impondrian el referendum, por tanto, si o si votarian los españoles.*

(60) [Economia #70]: ***HUELGA GENERAL SALVAJE***

9. Graus de toxicitat

A partir dels paràmetres i criteris comentats, podem concloure la següent gradació de la toxicitat, dividida en quatre nivells, i totes les característiques associades a cada grau (Taula 11). Cada característica queda classificada en un dels graus de toxicitat (del grau 1: No tòxic al grau 4: Molt tòxic).

No tòxic (GT1)	**Lleugerament tòxic (GT2)**	**Tòxic (GT3)**	**Molt tòxic (GT4)**
No presenta toxicitat	Podria ser considerat tòxic en alguns contextos, generalment, no té voluntat ofensiva.	És sarcàstic i/o presenta crítiques no constructives.	S'hi troben atacs personals, insults forts dirigits a

[18] Enciclopèdia.cat: https://www.enciclopedia.cat/ec-gec-0099407.xml

			persones concretes.
	Presenta paraules malsonants que no van dirigides a ningú en concret o expressen ràbia o frustració.	Fa burla o ridiculitza l'autor de contingut.	Presenta llenguatge molt ofensiu o abusiu.
	Conté sarcasme o ironia «blancs» (sense voluntat d'ofendre)	Discrepa agressivament o mostra agressivitat lleu.	Mostra una forta agressivitat.
	És una burla sense destinatari concret o que expressa frustració o ràbia.	Presenta bromes inapropiades.	És despectiu i/o degradant.
	És una mostra de supèrbia.	Ironia (quan el missatge real és el contrari del missatge literal) punyent o ofensiva.	Discurs intolerant sancionable: conté violència simbòlica o incita a la violència física; insulta directament als membres d'un col·lectiu determinat.
	Preguntes retòriques dirigides a un altre usuari.	Preguntes retòriques acompanyades de crítica mordaç.	
	Preguntes retòriques amb un to burleta.	Metàfores sarcàstiques, acompanyades de crítica.	
	Bromes acompanyades de sarcasme o que poden resultar ofensives per un col·lectiu.	Frases fetes o expressions comunes, quan contenen o van acompanyades de crítica o ràbia.	
	Atacs aparentment neutres (o amb un grau baix d'agressivitat però que es fan de manera reiterativa)	Atac o crítica contra un col·lectiu o un grup determinat: grup ètnic, racial, nacional, identitari, religiós, ideològic (inclosos els votants d'un partit polític), etc.	
	Paraules que no són insults directes	Atac a la intel·ligència i/o les habilitats lingüístiques de persones concretes.	

	però s'usen com a tal.	
	Discurs intolerant no sancionable: estereotips i prejudicis; missatges supremacistes o homonacionalistes; deshumanització, altredat o cosificació.	Comentaris humiliants o deshumanitzadors
		Preguntes retòriques amb un to burleta.
		Disfemismes
		Discurs intolerant no sancionable: estereotips i rumors, bocs expiatoris i amenaces; ridiculització per motius de pertinença a determinats col·lectius.

Taula 11: Graus de toxicitat

10. Correlacions entre trets:

Les correlacions que podem trobar entre els diferents paràmetres presentats ens poden ajudar a l'hora de classificar-los. Per exemple, que un comentari presenti insults (61), sempre implica que serà un comentari tòxic, però un comentari tòxic no sempre presentarà insults (62).

(61) [Política #62]: *¿Y no se será que más que que ellos sean unos genios, la base que les vota son directamente* ***subnormales****?* (Grau de Toxicitat = 4)

(62) [Política #16]: *Sánchez pacta lo que su amo pacte. No le han puesto ahí para pensar sino para obedecer y ya estuvo a punto de no poder pagar la hipoteca de su casa.* (GT = 3)

La presència d'insult implica que el grau de toxicitat serà un 3 o un 4.

(63) [Immigració #95]: *No es racismo,* ***progre****, es sentido común*

(64) [Política #98]: *Esta es la escoria política que tenemos (y hablo tanto de unos como de otros) porque es lo que muchos han votado (y lo peor es que seguirán haciéndolo). Y no hay alternativa que valga la pena, todos saben que el pueblo es una* ***masa de idiotas*** *y por ello hacen lo mismo*.

El llenguatge tòxic, el sarcasme i/o la ironia sempre implicaran un grau de toxicitat igual o superior a 2 (65).

(65) [Política #28]: *Y nos lo toma cuando le interesa. ¡que bonito simulacro de motín se montó para no tener que pagar el peaje de que el PSOE se abstuviese!* (GT = 2)

El tret binari d'argumentació sempre implica la presència de constructivitat (66), però un comentari constructiu no té per què presentar argumentació, ja que aquesta constructivitat pot ser deguda per altres causes, com el diàleg, que aporti noves perspectives o que aporti solucions (67).

(66) [Economia #24]: *Tienes razón, las pensiones son a título individual y así deben ser consideradas. Sin embargo, a menudo el punto de vista de las regiones se usa para enmierdar el debate, sobre todo cuando se sueltan datos sobre número de cotizantes frente a pensionistas, acusando sibilinamente a las zonas más pobres de vivir a cuenta de las más ricas. Sin embargo, se obvia que muchos de esos pensionistas actuales cotizaron toda su vida en esas regiones ricas (contando como en su momento activos),*

(67) [Economia #6]: *Para los bancos sí que hay enriquecimiento.*

Un comentari que s'hagi classificat amb el grau de toxicitat 4 no pot ser considerat constructiu (68), però un comentari amb un grau de toxicitat 3 o inferior sí que pot ser constructiu, per exemple si hi ha paraulotes, insults o menyspreu cap a l'autor de la notícia o d'un altre comentari però aporta informació nova, rellevant o fonamentada (69).

(68) [Economia #34]: *Jajajaja. Me rio yo del chupapollas de Rallo y sus seguidores* (No constructiu, GT=4)

(69) [Immigració #78]: *Y una puta mierda lo entendía así la izquierda antes. Quienes estamos en la CNT sabemos muy bien que las luchas entre trabajadores por el hecho de ser inmigrantes jamás formaron parte de ninguna estrategia ni táctica sindical de aquella época. Deja de mentir y de ensuciar el buen nombre de la clase trabajadora de entonces.* (Constructiu, GT=3)

3.4. Procés d'anotació

Un cop feta la primera aproximació al corpus i l'actualització de la guia d'anotació, es va començar amb el procés d'anotació de quatre temes del corpus NewsCom-HS: Economia, Immigració, Política i Religió. L'anotació del corpus es va dur a terme per part de tres anotadors diferents, dos estudiants del grau de lingüística i jo mateixa, i va consistir en l'anotació en paral·lel dels quatre temes seleccionats, és a dir, l'anotació independent per part dels tres anotadors un cop llegida la guia d'anotació i entesos els paràmetres i criteris d'anotació.

Per tal d'avaluar la qualitat de l'anotació, és a dir, la consistència de les dades anotades, s'ha realitzat una prova d'acord entre els anotadors. Aquesta prova permet avaluar la fiabilitat de l'anotació per part dels anotadors i saber si han aplicat els mateixos criteris. La prova d'acord entre els anotadors s'ha fet sobre el total de comentaris anotats, és a dir sobre els 1262 comentaris (309 d'economia, 239 de política, 298 de religió i 416 d'immigració). S'ha calculat l'*average pairwise percent agreement*, és a dir, la mitjana del percentatge d'acord entre parelles d'anotadors, i els resultats obtinguts, per tema i per tret anotat, són els que es presenten a la Taula 12:

Average pairwise percent agreement (%)	Constructivitat	Argumentació	Sarcasme	Burla	Insults	Llenguatge tòxic	Intolerància	Agressivitat	Toxicitat (SÍ/NO)	Toxicitat (grau)
Economia	89.644%	72.816%	88.565%	85.545%	92.880%	85.545%	96.8%	96.1%	81.446%	71.521%
Política	79.079%	70.711%	80.195%	78.522%	90.795%	84.379%	97.9%	100%	79.637%	64.714%
Immigració	70.032%	64.103%	85.577%	75.641%	86.538%	74.519%	76%	98.1%	74.679%	56.490%
Religió	76.812%	64.994%	82.163%	74.582%	92.419%	85.953%	84.3%	98.7%	73.467%	58.751%

Taula 12: Average pairwise percent agreement

Com podem veure a la Taula 12, els temes que presenten més acord en el grau de toxicitat són economia i política (amb un 71,52% i un 64,71% d'acord, respectivament), mentre que els temes amb més desacord són immigració i religió (amb un 56,49% i un 58,75% d'acord respectivament). Aquest tant per cent d'acord es podria donar pel fet que el tema d'immigració és el que més comentaris té anotats (416), és també el que té més comentaris tòxics, i perquè tant immigració com religió són els temes que poden presentar més comentaris polèmics, en què l'anotador té una implicació més emocional i que, per tant, poden ser més difícils de classificar.

Com es pot observar, hi ha molt menys desacord en el tret binari de toxicitat (SÍ/NO) que no pas en el grau de toxicitat. És a dir, que l'acord és molt més alt quan l'anotador només ha de dir si el comentari és tòxic o no que quan ha d'assignar un grau de toxicitat (l'acord puja un 10% en economia, un 14,92 % en política, un 14,71% en immigració i fins a un 18,18% en religió quan l'anotador ha de decidir si el comentari és tòxic o no). Això ens mostra que establir un grau de toxicitat és una tasca complexa, amb una part molt important de subjectivitat i que depèn molt de la interpretació de l'anotador.

Pel que fa a la resta de trets, els que presenten més acord són l'agressivitat, els insults i la intolerància, tot i que l'acord en aquest últim tret baixa al tema d'immigració, ja que és el tema on també trobem més comentaris amb contingut intolerant. Dels trets, el que presenta més desacord és l'argumentació. Aquest desacord es podria explicar pel fet que dos dels anotadors van considerar argumentatiu també els comentaris que fomentaven el diàleg, mentre que el tercer anotador no ho va considerar així. Per tant, aquest seria un cas en què s'hauria de modificar la guia d'anotació perquè quedés clar què es considera argumentatiu.

A partir dels resultats obtinguts en la prova d'acord entre els anotadors, hem procedit a l'elaboració del *Gold Standard*, és a dir, la versió del corpus que servirà de model i el que s'utilitzarà per entrenar i avaluar els sistemes d'aprenentatge automàtic. Per obtenir aquest *Gold Standard* hem classificat els comentaris en tres grups tenint en compte l'anotació dels graus de toxicitat: els acords totals, que són els comentaris en què els tres anotadors estan d'acord en l'anotació del grau de toxicitat; els acords parcials, que són els comentaris en què dos anotadors estan d'acord i un no; i els desacords totals, que inclou els comentaris en què cap dels anotadors coincideix en l'anotació. Els comentaris amb acord total i els comentaris amb acord parcial, on ens quedarem amb l'anotació de la majoria, passen a formar part del *Gold Standard*. En canvi, els desacords totals són revisats per cinc anotadors (els tres anotadors del corpus i dos anotadors sènior) per tal de discutir i acordar l'anotació final. Un cop consensuada l'anotació, és a dir, el grau de toxicitat, s'inclouen també en el *Gold Standard*. Pel que fa a la resta de trets, com que es tracta de trets binaris s'ha pres l'anotació majoritària com la definitiva.

La Taula 13 mostra els diferents tipus de desacords totals que podem trobar, tant per tema com en total:

	1/2/3	**1/2/4**	**1/3/4**	**2/3/4**	**Total desacords**
Economia	16	1	0	4	21
Immigració	24	9	4	14	51
Política	24	1	0	2	27

Religió	17	3	2	12	34
Total	81	14	6	32	133
Percentatges totals	60.45%	10.45%	4.45%	23.88%	

Taula 13: Desacords totals.

Com podem veure a la Taula 13, els desacords més comuns són els que un anotador adjudica el grau de toxicitat GT1, un altre el grau de toxicitat GT2 i l'últim anotador un grau de toxicitat GT3, amb un 60.45% del total de desacords en els quatre temes. El següent desacord més comú és el que un anotador adjudica el grau de toxicitat GT2, l'altre el grau GT3 i l'últim el grau GT4, amb un 23.88% del total de desacords. També com podem veure, el tema que presenta més desacords és el tema de la immigració, amb 51 comentaris en desacord total, seguit pel tema de religió, amb 34 comentaris en desacord total. També es tracta dels temes que tenen més comentaris i els que tenen més comentari tòxics, especialment el d'immigració (veure Taula 14).

Per exemple, un dels comentaris que va donar desacord total i que vam haver de revisar amb els cinc anotadors va ser el següent:

(70) Política #93: *Te falto el ironic?*

En el comentari (70) vam tenir un anotador que li va donar un grau de toxicitat GT2, un altre un grau de toxicitat GT1 i l'últim un grau de toxicitat GT3. Aquest desacord es va produir per la subjectivitat que implica el sarcasme o la ironia. L'anotador que li va donar un grau de toxicitat GT3 ho va fer perquè el va considerar sarcàstic, i això implica un nivell de toxicitat alt (GT3), mentre que l'anotador que va donar un grau de toxicitat GT2 el va considerar que es tractava de *sarcasme blanc*, és a dir, un sarcasme sense intenció d'ofendre i, per tant, amb un nivell de toxicitat menor (GT2); i l'anotador que li va donar un grau de toxicitat GT1 no va considerar que el comentari presentés sarcasme de cap tipus. A la reunió amb els cinc anotadors es va acabar adjudicant un grau de toxicitat GT1, per les mateixes raons que l'últim anotador.

També vam trobar desacord total en el següent comentari (71):

(72) Política #19: *ese no está en el TS?*

Un anotador li va donar un grau de toxicitat GT1, un altre GT2 i l'últim un grau de toxicitat GT3. En aquest cas, el desacord es va donar perquè dos dels anotadors van tenir en compte el context, és a dir, el fil de la conversa perquè es tractava de comentaris que es presentaven de manera consecutiva (71-73):

(71) [Política #18]: *Como no les ha gustado a los indepes Iceta y probablemente tampoco estos dos, yo propondría a Tomás de Torquemada, un hombre conocido por su tolerancia. Seguramente será del agrado de los indepes.*

(72) [Política #19]: *ese no está en el TS?*

(73) [Política #20]: *No, está en Waterloo*

Els anotadors no van aplicar el criteri que es definia a la guia on s'establia que no es tindria en compte el context (és a dir, els comentaris que es contesten entre ells) a l'hora d'anotar la toxicitat i d'aquí va sorgir la disparitat en l'anotació. En les revisions amb els cinc

anotadors es va acordar que el context no es tindria en compte, tal i com s'explica a la guia d'anotació, i és va decidir finalment adjudicar-li un grau de toxicitat GT1.

Un exemple de desacord total 2/3/4 seria el següent:

(74) [Economia #93]: *La capitalización es un sistema egoista e insolidario que se corresponde con los anglosajones, que son insolidarios e individualistas, que no creen en los lazos familiares y en la solidaridad entre los ciudadanos de una misma la sociedad, sino que creen en la ley de la selva y si te jodes es culpa tuya y de nadie mas. No gracias.*

Aquest comentari (74) va ser marcat com a tòxic per tots els anotadors, però amb diferent grau de toxicitat: GT2 per part d'un anotador i GT3 i GT4 per part dels altres dos. L'anotador que va marcar un grau de toxicitat GT2 ho va fer pel llenguatge tòxic ("*te jodes"*) i perquè va considerar que "*que son insolidarios e individualistas, que no creen en los lazos familiares y en la solidaridad entre los ciudadanos de una misma la sociedad"* no eren insults directes. En canvi, l'anotador que va adjudicar un grau de toxicitat GT3 ho va fer perquè va considerar-ho un atac a una nacionalitat i, per tant, intolerància; mentre que l'anotador que va marcar el grau de toxicitat GT4 ho va fer perquè va considerar que els insults eren directes i, per tant, era necessari un grau de toxicitat major (GT4). En la revisió amb els cinc anotadors vam adjudicar-li un grau de toxicitat GT2, per la presència de llenguatge tòxic i perquè vam considerar que els insults no eren directes ni massa ofensius.

Com podem veure, no només el sarcasme o el context poden produir interpretacions subjectives diferents, sinó que fins i tot altres trets com els insults, que en la majoria de temes presenta un acord per sobre del 90%, poden presentar-nos dubtes i diferències entre anotadors. Això ens mostra la complexitat de la tasca d'anotació, que està molt condicionada per la subjectivitat i, en especial, dels graus de toxicitat.

A partir d'aquests acords totals i parcials i els desacords revisats s'ha creat el *Gold Standard,* és a dir, un corpus anotat amb una anotació fiable i de qualitat. El *Gold Standard* serà el que s'utilitzarà per entrenar sistemes de detecció automàtica de la toxicitat basats en aprenentatge automàtic.

La Taula 14 presenta els comentaris del *Gold Standard*, classificats per tema i per grau de toxicitat:

Comentaris	**Total**	**No tòxics GT1**	**Tòxics**	**GT2**	**GT3**	**GT4**
Economia	309	209	100	70	21	10
Immigració	416	221	195	109	62	22
Política	239	129	110	56	32	19
Religió	298	179	119	67	29	18
Totals	1262	738	524	302	144	69

Taula 14: La toxicitat en el Gold Standard

La Taula 15 mostra els percentatges de comentaris tòxics per tema i els percentatges de comentaris de cada grau de toxicitat (GT2, GT3 i GT4) dins dels comentaris tòxics:

	% de comentaris tòxics	% de comentaris GT2	% de comentaris GT3	% de comentaris GT4
Economia	32.36%	70%	21%	10%
Immigració	46,88%	55,90%	31,79%	11,28%
Política	46,03%	50,91%	29,09%	17,27%
Religió	39,93%	56,30%	24,37%	15,13%
Total	41,52%	57,63%	27,48%	13,17%

Taula 15: Percentatges de comentaris tòxics al Gold Standard

Com podem veure a la Taula 15, els temes que presenten un percentatge més alt de comentaris tòxics són immigració i política, amb un 46,88% i un 46,03% respectivament, mentre que el tema que presenta menys comentaris tòxics és economia, amb un 32.36% de comentaris tòxics. Com ja hem comentat, és esperable que immigració sigui un dels temes més tòxics, però també el de política, ja que es tracta de temes que presenten molta polèmica. En canvi, sembla que en el cas d'immigració hi ha més toxicitat de nivell GT2, mentre que en el tema de política, tot i que el número de comentaris és inferior, hi ha més toxicitat i és el que presenta graus de toxicitat més elevats (17,27% en GT4), ja que l'insult, la burla i ridiculització a partits polítics o a polítics és més freqüent.

Els comentaris tòxics més freqüents són els comentaris amb grau de toxicitat GT2 en tots els temes, especialment en el tema d'economia, on són el 70% del total de comentaris tòxics, i on probablement es fa més ús de la ironia i el sarcasme blanc.

4. Conclusions

En aquest treball s'ha presentat, primer, una revisió dels treballs anteriors sobre el discurs d'odi i la seva classificació que ens ha ajudat a veure on es situa l'estudi del llenguatge d'odi o tòxic, ens hem centrat especialment en el tipus d'informació que inclouen els corpus anotats amb aquest tipus d'informació en el marc del processament del llenguatge natural, que ens ha proporcionat uns primers paràmetres amb els quals treballar. Segon, s'ha proposat una guia d'anotació amb nous trets i criteris que ens ajudin a descriure la toxicitat, i a classificar-la en funció del grau de toxicitat que presenten els comentaris. Per últim, s'ha presentat els resultats que hem obtingut de l'anotació del corpus NewsCom-HS a partir de la nostra guia d'anotació i la creació del *Gold Standard*. Aquest *Gold Standard* presenta un total de 1262 comentaris, 524 dels quals són tòxics. D'aquests 524 comentaris, un 57,63% tenen un grau de toxicitat GT2, un 27,48% presenten un grau de toxicitat GT3 i un 13,17% tenen un grau de toxicitat GT4.

A partir dels resultats obtinguts, hem pogut comprovar que alguns trets ens ajuden més que altres a l'hora d'identificar i classificar la toxicitat, com poden ser els insults, la intolerància o l'agressivitat, mentre que altres, com l'argumentació o la constructivitat són menys determinants.

Aquest treball m'ha suposat un primer acostament al món de la investigació que m'ha permès aprendre i conèixer no només tot el procés que implica la investigació, sinó també tot l'aprenentatge que suposa el procés d'anotació d'un corpus. En aquest aprenentatge s'inclou la creació de la guia d'anotació, el procés d'anotació amb diferents anotadors, i les proves per calcular l'acord entre els anotadors i la creació del *Gold Standard*.

Aquest treball forma part d'un projecte que encara s'està desenvolupant, aquí només presentem l'anotació de quatre temes dels nou temes que té el corpus NewsCom-HS. Per tant, aquest treball ha suposat una primera aproximació al problema que és la identificació i proposta d'anotació de la toxicitat en el llenguatge.

5. Bibliografia

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Sanguinetti, M. (2019). *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 512-515.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *CEUR Workshop Proceedings*, *1816*, 86-95.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4).

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2020). The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments. *Corpus Pragmatics*, *4*(2), 155-190.

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, *14*(8), 1-16.

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018*, 61-66.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *25th International World Wide Web Conference, WWW 2016*, 145-153.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., & Bosco, C. (2017). Hate speech annotation: Analysis of an Italian twitter corpus. *CEUR Workshop Proceedings*, *2006*.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis.

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing, (2012), 1-10.

Taulé, M., Nofre, M., González, M., & Martí, M. A. (2019). Focus of negation: its identification in Spanish.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceeding LSM '12 Proceedings of the Second Workshop on Language in Social Media*, (Lsm), 19-26.

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter, 88-93.

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks, 78-84.

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018). Inducing a lexicon of abusive words ? a feature-based approach. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 1046-1056.

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia i Comunicació

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

Tel. +34 934 035 594
fil-coord@ub.edu
www.ub.edu

Declaració d'autoria

Amb aquest escrit declaro que soc l'autor/autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18 del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 12 de juny de 2020

Signatura: [redacted]