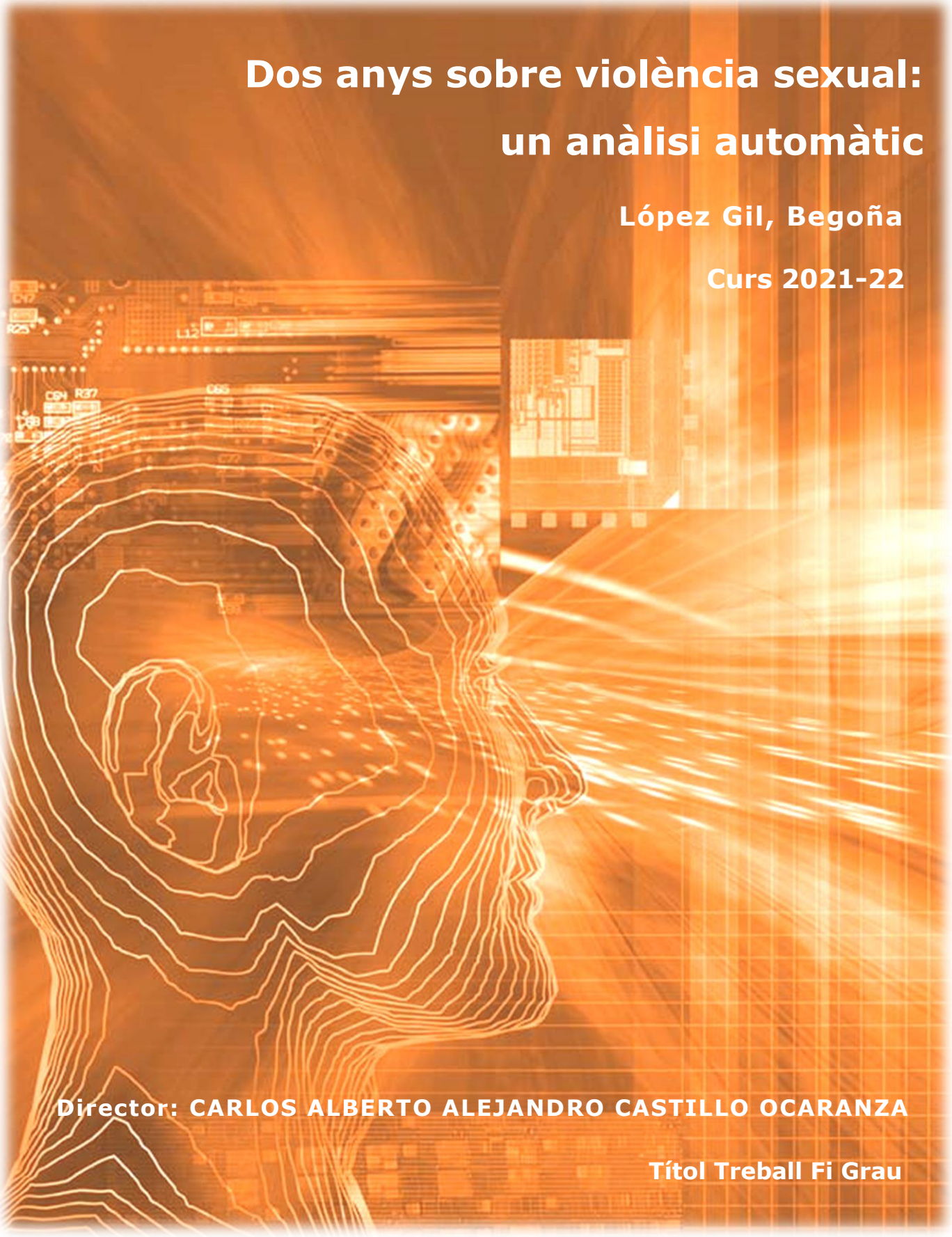


Dos anys sobre violència sexual: un anàlisi automàtic

López Gil, Begoña

Curs 2021-22

Director: CARLOS ALBERTO ALEJANDRO CASTILLO OCARANZA

Títol Treball Fi Grau

upf. Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Treball de Fi de Grau

Dos anys de notícies de violència sexual: un anàlisi automàtic

TREBALL FI DE GRAU DE

Begoña López Gil

Director: Carlos Castillo

Grau en Enginyeria Matemàtica en Ciència de Dades

Curs 2021-2022

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Agraïments

En primer lloc, agrair al director d'aquest treball al Carlos Castillo qui ha sigut clau, per tal d'encaminar i acompanyar-me aportant els seus coneixements. No hagués estat possible aquest treball sense ell.

Segonament, vull donar les gràcies a la Marilena Budan, la persona que va fer el projecte del qual sorgeix aquest treball. Gràcies per la teva aportació i per haver participat donant el teu punt de vista durant tot el procés.

En Fedor Vitugin per haver dedicat el seu temps i haver-me proporcionat la seva ajuda per poder recol·lectar les dades.

A la meva família i a la meva amiga Belén, per tot el suport emocional. Gràcies per haver-me recolzat durant tota la meva vida acadèmica.

Per últim, a tots els investigadors per l'aportació científica sobre la qual es sustenta aquest treball.

Índex

1. INTRODUCCIÓ

1.1 Context

Gairebé una de cada 3 dones al món ha patit violència sexual en algun moment de la seva vida. L'anàlisi de l'Organització Mundial de la Salut realitzat entre 2000 i 2018 a 161 països va determinar que un 30% de les dones havien patit en alguna ocasió violència sexual. És considerada per l'OMS un greu problema de salut pública (World Health Organization, 2021). A Espanya, les dades publicades el maig de 2022 per l'Institut Nacional d'Estadística indiquen que 30.141 dones van ser víctimes de violència de gènere el 2021 xifra que ha augmentat un 3,2% respecte al 2020 on el nombre de víctimes eren 29.215 (Instituto Nacional Estadístico, 2021).

Amb freqüència podem veure biaixos i males pràctiques als mitjans de comunicació reproduint el tractament de la informació de la violència sexual com els estereotips sexistes de víctimes i agressors, la descontextualització, la difusió de mites i la culpabilització de víctimes (González, G. i Jiménez, M.L, 2018) Aquestes males pràctiques es poden veure no només en el contingut dels articles sinó també de manera implícita en el llenguatge utilitzat (O'Hara S, 2012).

Segons l'últim Estudi Anual de Xarxes Socials 2021 realitzat per IAB Spain el 85.5% de persones que utilitzen internet fan ús d'alguna xarxa social, el que representa 26,6 milions de persones d'entre 16 i 70 anys. Dels usuaris enquestats, un 31% afirmaven que segueixen a mitjans de comunicació a les xarxes socials (IAB Spain, 2021). Està clar que la presència de les xarxes socials han suposat un canvi en el procés de comunicació. i s'han convertit en una eina de creació de contingut i, sobretot, de la seva posterior difusió. El periodisme actual no només necessita contar històries i fer-ho de la millor manera sinó que necessita saber on i a qui contar-les. És per això que Twitter s'ha convertit en un sistema d'alarma informativa per la seva capacitat de detectar, de forma instantània notícies importants i d'interès de qualsevol lloc i temps (Martínez, H. i Sánchez, G, 2022). És per això que la xarxa social Twitter s'ha convertit en un element imprescindible a la premsa.

1.2 Plantejament del problema

La violència sexual és un greu problema de salut pública, però, ens arriben realment aquestes notícies? Influeixen els mitjans de comunicació en la perpetuació dels estereotips i mites? Durant quant de temps es parla d'un cas als medis de comunicació?

1.3 Objectius

Partint del treball de fi de grau 'Analysis of online news articles: the coverage of sexual violence in Spanish media' de Marilena Budan, l'objectiu principal es aprofundir en la seva recerca ampliant el component temporal durant el procés.

Els objectius, igual que en l'anterior treball, es poden desglosar en tres apartats:

A) Col·lecció de dades:

El primer repte és ampliar considerablement el conjunt de dades anterior. Tant en número de medis com en quantitat per medi, recollint els articles publicats durant un període de temps determinat.

B) Agrupació de casos:

En segon lloc agrupar els articles a través d'un model automatitzat segons el cas que tracten tenint en compte en quina.

C) Anàlisis d'articles i casos:

Finalment, l'últim objectiu és analitzar els articles i els casos afegint el component longitudinal. Quant de temps dura un cas? Es parla més d'un tipus de casos que d'altres? Es parla igual dels casos que duren més temps que dels que es parla poc?

1.4 Estructura del document

Aquest treball es divideix en 6 capítols:

- Capítol 1 *INTRODUCCIÓ* descriu el context del problema i especifica els objectius del projecte.

- Capítol 2 *ESTUDI PREVI* s'explica la recerca prèvia a la realització del projecte.

- Capítol 3 *METODOLOGIA* on s'exposen els passos seguits per complir amb els objectius detallats anteriorment.

- Capítol 4 *RESULTATS* es detallen els resultats obtinguts un cop seguits els passos explicats anteriorment.

- Capítol 5 *DISCUSSIÓ* i *CONCLUSIONS* desenvolupa els resultats detallats anteriorment utilitzant la recerca explicada al segon capítol.

2. ESTUDI PREVI

Aquest treball de fi de grau parteix d'un anterior realitzat al 2020-2021 per Marilena Budan anomenat 'Analysis of online news articles: the coverage of sexual violence in Spanish media' (Budan, M., Castillo, C., 2022). L'objectiu d'aquest capítol és resumir aquest treball per tal d'entendre les bases d'aquest.

2.1. Objectius

Aquesta recerca utilitza articles de noticies escrits per a un públic ampli i publicats per grans organitzacions de mitjans de comunicació d'Espanya. Les aportacions de la investigació inclouen:

1. un classificador automatitzat per trobar articles de notícies en castellà sobre violència sexual,
2. una col·lecció d'URL que apunten a articles de notícies sobre violència secual publicats per grans mitjans de comunicació a Espanya,
3. una metodologia per agrupar aquests articles de notícies en grups d'articles que probablement facin referència al mateix esdeveniment,
4. una col·lecció d'expressions regulars per inferir descripcions rellevants dels articles de notícies, i
5. una anàlisi en la qual comparem la prevalença de diferents descriptors amb les estadístiques oficials sobre violència sexual.

2.2. Col·lecció de dades

Es va començar amb una llista de les 15 fonts informatives generalistes a Espanya amb més audiència, segons el Digital News Report de Reuters Institute (Newman, N. et al. 2020). Tots aquests utilitzen llocs de xarxes socials, com Twitter i Facebook, per difondre el seu contingut amb l'objectiu d'augmentar el trànsit als seus llocs web (Ahmad, A. N. 2010).

Havent determinat quin és el compte més adequat es van recollir els 10.000 tuits més recents al setembre de 2020 utilitzant l'API pública que ofereix Twitter. Aquest tuits es van classificar mitjançant un classificador supervisat basat en la regressió logística. Aquest va ser entrenat de manera iterativa, començant amb un classificador creat etiquetant manualment els 2.000 tuits més recents de cada font.

Després, aquest classificador es va aplicar a tot el corpus de tuits descarregats i s'utilitza per guiar un etiquetatge més gran. El classificador final es va fer al llarg de tot el conjunt de dades, obtenint una precisió del 85% i un reclam de 0,83.

A continuació, s'ha considerat positivament totes les URL contingudes als tuits classificats. Dues de les 15 comptes originals (El País i El Diario) van utilitzar una tècnica per amagar les URL dels tuits, i per tant no es van tenir en compte.

Es van seguir les URL descarregant els enllaços de destinació, cosa que va ser possible en el 50-60% dels casos. Els articles es van convertir en articles descarregats en format

NewsM-G2, una forma estandarditzada d'emmagatzemar articles en format XML definit per l'International Press Telecomunications Council (IPTC)[1]. El conjunt de dades final conté 496 articles de notícies.

2.3. Agrupació de notícies

És comú que diferents articles parlen d'un mateix cas, per tant es va decidir agrupar els articles descarregats anteriorment que parlen d'un mateix esdeveniment. Per fer-ho primer es construeix una funció de distància supervisada. Aquesta retorna la probabilitat de que dos articles facin referència al mateix esdeveniment.

La probabilitat es calcula mitjançant un classificador de regressió logística calibra entrenat en una mostra de parells d'articles etiquetats manualment. El classificador resultant té una precisió del 94% i un reclam del 90%.

L'agrupació es va crear mitjançant *Agglomerative Jerarquical Clustering*, que agrupa objectes per formar un arbre binari amb la modalitat d'enllaç 'mitjana'.

S'han categoritzat els casos coberts per només una font com a cobertura baixam les cobertes per dues o tres fonts amb cobertura mitjana i casos coberts amb més de tres fonts com a cobertura alta.

2.4. Anàlisis automatitzat

Es va avaluar el contingut dels articles de notícies en dos nivells: per cluster i per article.

Primerament l'anàlisis per cluster es busquen discrepàncies entre les distribucions de categories o descriptors en els casos, i les mateixes distribucions en les enquestes realitzades pel Govern d'Espanya.

Per l'anàlisis per article s'han considerat la presència o absència de tipus específics d'informació en diferents parts d'articles de notícies.

[1] https://iptc.org

3. METODOLOGIA

3.1. Procediment

El procediment es divideix en tres parts:

A) Col·lecció de dades.

En aquesta primera part es recol·lecten tots els tuits des del gener de 2020 fins l'abril 2022 de 24 medis de comunicació espanyols més populars. A continuació es considera el text d'aquests tuits per classificar-los segons si són de violència sexual o no.

Un cop tenim els que són de violència sexual és recol·lecten els articles seguint l'enllaç que hi ha als tuits. Tant els tuits com els articles es recullen tenint en compte que es tracta d'un tema sensible i s'han de fer anònimes les dades.

B) Classificació de casos

Tot seguit, s'agrupen els articles per casos. Tenint en compte que s'han recollit dades de 28 mesos i s'ha de tenir en compte el temps.

C) Anàlisis d'articles

L'anàlisi es pot dividir en dos subapartats: anàlisi de casos i d'escriptura d'articles. Per extreure informació sobre les característiques més comunes.

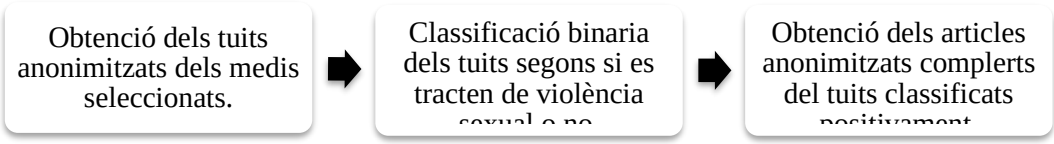
3.2. Col·lecció de dades

S'han escollit els medis de comunicació amb més audiència online d'acord amb el *Digital News Report Espanya 2021* (Amoedo, A., 2021). Aquest informe presenta un llistat dels 39 medis de comunicació amb més audiència setmanal online alguns d'ells sense compte de Twitter, de temàtiques concretes (d'economia o esport) i regionals.

Aleshores d'aquesta llista s'han seleccionat els medis nacionals d'informació general que tenen un compte a Twitter actiu, quedant una llista final de 24 medis: 20 Minutos, El País, La Sexta, El Confidencial, RTVE, La Vanguardia, Diario ABC, La SER, Ok Diario, El Diario, El Periódico, La Razón, Informativos Telecinco, El Mundo, Antena 3 Noticias, Noticias Cuatro, COPE, Onda Cero, El Público, El Huff Post, El Español, Libertad Digital, Europa Press, i Vozpópuli.

Els passos seguits per aconseguir el conjunt de dades són:

Figura 1. Metodologia col·lecció de dades



El número d'usuaris que consulten les notícies a través de xarxes socials com Facebook i Twitter augmenten. Conseqüentment, els medis de comunicació incorporen cada dia més les xarxes socials, convertint-se així Twitter una part important al procés de comunicació (Martínez, H. i Sánchez, G, 2022).

Per això que Twitter és la font inicial de dades amb la fi d'extreure el contingut de les notícies que hi ha incorporades als tuits. Primerament es recol·lecten els tuits per poder classificar-los i un com classificats s'extreu el contingut de les notícies que hi ha enllaçades.

3.2.1. Obtenció dels tuits

El primer pas per recol·lectar els tuits és identificar els comptes principals a Espanya de cadascun dels mitjans de comunicació, veure *ANNEX 2. Col·lecció de dades* per consultar el comte dels mitjans enumerats anteriorment amb les seves descripcions.

Un cop seleccionats els comptes del 24 mitjans de comunicació recollim els tuits des de gener 2020 fins abril 2022. Per fer-ho he utilitzat un programa que recol·lecta aquest tuits mes a mes, veure *ANNEX 1. Dades i codi.*

Per poder fer aquesta recol·lecció he utilitzat l'accés acadèmic, d'aquesta manera puc accedir a tuits sense una limitació de temps i una limitació de quantitat de 10M al mes. A més amb aquest tipus d'accés puc obtenir de manera directa els enllaços de les noticies referenciades als tuits igual que a la resta d'estadístiques.

Degut a la sensibilitat del tema que ens ocupa, la violència sexual, a l'hora de descarregar les dades prèviament les fem anònimes. Per fer-ho utilitzarem *scrubadub,* en concret el detector ja entrenat *TextBlobNameDetector*, aquesta llibreria substitueix els noms propis pel terme *'[NAME]'*. Encara que és poc precisa degut a la quantitat de falsos positius que es poden apreciar és la millor solució ja que és eficient. Aleshores els texts resultants no tindran cap tipus de dada personal com noms, organitzacions o països.

Finalment, s'obtenen els tuits guardats en dos formats diferents per a cada dia i medi, com a *jsonlines* y com *csv*. Per cada tweet es guarda el identificador, el identificador de l'autor, el text, la data de creació, si és contingut sensible, el número de retuits, respostes, m'agrades, cites i mencions.

3.2.2. Classificació dels tuits

L'objectiu d'aquesta part és classificar els tuits descarregats anteriorment utilitzant el seu text per tal de determinar quins tracten casos de violència sexual ja que aquesta és la manera més òptima de fer-ho (Budan, M., Castillo, C., 2022).

Preparació de les dades

El primer pas per poder classificar els tuits utilitzant algoritmes d'aprenentatge automàtic és transformar el text en vectors numèrics. Però abans aquest text s'ha de normalitzar: posar en minúscules el text, eliminar signes de puntuació, símbols i les paraules buides.

Al conjunt de paraules buides afegirem el terme *name* degut a que quan es fa anònim el text a l'apartat *3.2.1. Obtenció dels tuits* es substitueixen els noms propis per '*[NAME]*', aleshores aquesta paraula no ens aporta cap informació durant la classificació i ens podria esbiaixar els resultats.

Un cop els tuits estan normalitzats es transformen en vectors numèrics. Per fer-ho s'ha utilitzat el mètode count vectorizer. Aquest crea una matriu de la mida *NxM* on *N* és el número de mostres i *M* el número de termes únics al conjunt de dades. Cada posició *[n,m]* especifica el número de cops que apareix a la mostra *n* el terme *m*.

Selecció del model

Utilitzem un model de regressió logística supervisada el qual es va seleccionar a l'anterior treball - veure capítol *2. ESTUDI PREVI* - ja que és un model que s'havia mostrat anteriorment que dóna bons resultats per fer anàlisis de text amb notícies. Aquest model prediu la probabilitat de pertànyer a una classe donades diferents variables aplicant la funció sigmoide *(1)* a una combinació lineal de variables. (Harrell, F. E. 2015)

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad (1)$$

Aquesta funció retorna la probabilitat de pertànyer a la classe etiquetada com a 1. Per obtenir la classificació s'utilitza un llindar on els que són és grans s'etiqueten com a 1, els que són més petits com a 0.

Aquest llindar es sol escollir com a 0.5, però també es pot variar per obtenir resultats més precisos. La manera de fer-ho més òptima en un classificador binari és maximitzant la *F1-score (2)*, la mitjana harmònica de la precisió i el reclam .

$$F_1 = 2\frac{precisió \cdot reclam}{precisió + reclam} \qquad (2)$$

Aquesta mesura s'utilitza habitualment per mesurar com de bé funciona un classificador binari quan una de les mostres es considera rara.

Etiquetar dades

Per poder utilitzar un model de classificació supervisat hem d'entrenar-ho amb un conjunt de dades prèviament etiquetades. Per fer-ho aprofitarem les dades que va classificar Marilena Budan - veure amb més detall com es van classificat al *2. ESTUDI PREVI*. Aquestes dades estan etiquetades segons el codi de la taula següent:

Taula 1. Valors/significat classificació tuits

Valor	Singnificat
0	No són sobre violència sexual.
1	Són sobre violència sexual.

El conjunt de dades etiquetat es troba molt desbalancejat, el primer pas és balancejar-ho ja que quan tenim una classe no balancejada es produeix cert esbiaix cap a la classe majoritària. Aleshores utilitzant la tècnica d'*undersampling*[2] (Susan, S. Kumar, A. 2020). Eliminem mostres etiquetades com a 0 de manera aleatòria per tal d'aconseguir un conjunt de dades amb un 75% de la mostra classificats negativament.

Per poder millorar el conjunt d'entrenament s'han realitzat els següents passos:

i. Executar el classificador utilitzant el nou conjunt de dades.

ii. Veure on se situa el threshold més òptim amb l'AUC.

iii. Afegir tuits que es troben a prop d'aquest límit al conjunt d'entrenament mantenint el balanceig.

iv. Tornar al pas i de manera cíclica amb el nou subconjunt.

Aquests passos els podrem seguir fins que els resultats del nostre classificador siguin suficientment bons. És importar al pas *iii* detectar quins tuits classifica malament i si són d'algun tema en específic per tal d'afegir contra exemples.

Classificació final

Després d'haver millorat el conjunt d'entrenament i tindre un conjunt de mostres suficientment gran com per obtenir bons resultats durant la classificació podem utilitzar-ho per classificar la resta de mostres utilitzant el llindar que maximitza la *F1-score*.

Degut la gran quantitat de dades que hi ha no és viable comprobar manualment els resultats. Per això és importat tindre un model de classificació entrenat ho millor possible ja que retornarà la classificació final.

[2] Estratègia que consisteix en eliminar mostres de la classe majoritària. Hi ha diferents criteris per decidir quines mostres s'han d'eliminar per tal d'evitar perdre informació important.

3.2.3. Obtenció d'articles

L'últim pas per a la preparació de dades és recollir els articles de les noticies. Per fer-ho, utilitzarem els links que obtenim al recollir els tuits.

Format

Seguirem el format NewsML-G2 escollit anteriorment per Marilena Budan -veure capítol 2-. Un format estàndard per guardar articles com XML definit per la International Press Telecommunications Council (IPTC).

Taula 2. Estructura NewsML-G2 utilitzada per guardar articles de notícies XML.

```
<newsItem standard="NewsML-G2" guid=[ARTICLE'S IDENTIFIER] version="1"
conformance="power" standardversion="2.15">
    <catalogReg href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-Standards_22.xml"/>
    <itemMeta>
        <itemClass qcode="ninat:text"/>
        <provider qcode="ninat:[MEDIA]"/>
        <itemMeta>[EXTRACTION DATETIME]</itemMeta>
        <pubStatus qcode="stat:usable"/>
        <contributor>
            <name>Twitter</name>
            <tweet_id>[TWEET ID]</tweet_id>
        </contributor>
    </itemMeta>
    <contentMeta>
        <contentCreated>[ARTICLE'S PUBLICATION DATETIME]</contentCreated>
        <located type="cptype:country" qcode="iso3166-1a2:ES">
            <name>[PUBLICATION'S COUNTRY]</name>
        </located>
        <creator>
            <name>[ARTICLE'S AUTHOR]</name>
        </creator>
        <headline xml_lang="es">ES</headline>
        <infoSource uri=[ARTICLE'S CANONICAL URL]</>
    </contentMeta>
    <groupSet root="G1">
        <grup id="G1" role="group:main">
            <itemRef residref="[ARTICLE'S IDENTIFIER]:headline">
                <itemClass qcode="ninat:text"/>
                <provider qcode="ninat:[MEDIA]"/>
                <pubStatus qcode="stat:usable"/>
                <title>[ARTICLE'S TITLE]</title>
                <description role="drol:headline">[ARTICLE'S
                SUMMARY]</description>
            </itemRef>
            <itemRef residref="[ARTICLE'S IDENTIFIER]:article">
                <itemClass qcode="ninat:text"/>
                <provider qcode="ninat:[MEDIA]"/>
                <pubStatus qcode="stat:usable"/>
                <description role="drol:article">[ARTICLE'S TEXT]
                </description>
            </itemRef>
        </grup>
    </groupSet>
</newsItem>
```

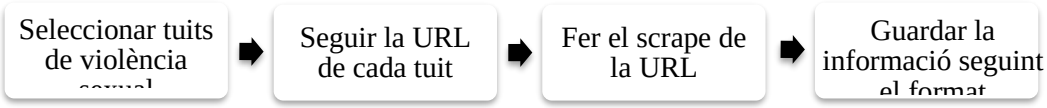
Tal i com es pot veure a la taula anterior, les dades guardades es poden assignar en diferents categories:

- Identificador de d'article: compost a partir de "urn:[MEDIA]:[EXTRACTION DATE]:[TWEET ID]"
- Dades d'extracció i de la font: especificar la data d'extracció i l'identificador del tuit.
- Informació de l'article: font d'URL, DateTime de la publicació, la font que el va publicar, el país i l'autor.
- Encapçalament de l'article: títol i subtítol.
- Text de l'article: el cos.

Procès per obtenir articles

Per obtenir els articles de les notícies es segueix el següent procés:

Figura x. Metodologia obtenir articles



Seguint una metodologia molt similar a l'anterior treball, el primer pas és seleccionar els tuits de violència sexual classificats a la secció 3.2, aleshores seguirem la URL de cada tuit emmagatzemada al fitxer tipus JSON creat a la secció 3.2.

Aleshores, farem el *scrape* i guardarem la informació en el format indicat prèviament. La informació sobre el contingut de la notícia, títol, subtitol, i cos la guardarem de manera anònima, igual que quan hem recol·lectat els tuits al apartat 3.1.

A diferència del treball anterior, on hi havia limitacions a l'hora de seguir l'URL dels tuits, ara tenim l'avantatge de tenir les credencials d'acadèmic de Twitter. Aleshores quan es recol·lecten els tuits a la secció 3.2.1 obtenim directament els enllaços als articles. Encara així perdem una petita quantitat ja que hi ha tuits sense enllaços o enllaçats a webs que no són la seva principal.

3.3. Classificació per casos

Seguint el procediment de l'anterior estudi un cop tenim tots els articles el següent pas és classificar els articles per casos. Ja que diferents medis pengen habitualment articles d'un mateix cas. És importat considerar que com tots els articles són de violència sexual el vocabulari utilitzat és semblant i classificar per casos no és trivial.

En aquest punt és importat tindre en compre el gran volum d'articles que estem utilitzant i que per tant s'ha de fer de la manera més eficient possible.

La metodologia seguida en aquest apartat és diferent que a l'anterior estudi, ja que els seus resultats són precisos però no és un procés òptim. Primer, seleccionem amb quina representació del text (titol, subtitol i cos) mesurarem la semblança dels documents.

3.3.1. Representació del text

El mètode utilitzat per la representació del text és el *Term Frequency Inverse Document Frequency (TF-IDF)*. Ja que a l'anterior estudi es va demostrar que era l'atribut més important per detectar parells d'articles sobre el mateix cas - veure capítol 2.

El TF-IDF és una representació basada en l'espai vectorial, és una tècnica comuna utilitzada per el processament de text. En aquesta representació, el terme freqüència (TF) per a cada paraula es multiplica per la freqüència inversa del document (IDF). Normalitza i redueix el pes de cadascun dels termes que apareixen amb freqüència a la col·lecció, d'aquesta manera redueix la importància dels termes comuns a la col·lecció. Assegurant que la concordança dels documents es veu més influenciada per la presencia de paraules que tenen freqüències més baixes. (Patil, L. H., Atique, M. 2013)

Donat un terme t i un document d el TF-IDF s'obté amb l'expressió *(3)*, on:

- N és el nombre total de documents.
- $tf_{t,d}$ és el nombre de cops que apareix el terme t al document d.
- df_t és el nombre de documents que contenen el terme t.

$$TF - IDF_{t,d} = tf_{t,d} \cdot idf_t = tf_{t,d} \cdot log(\frac{N}{df_t}) \quad (3)$$

3.3.2. Mètriques de semblança

Donats dos elements representats com a vectors, es pot mesurar la seva semblança utilitzant la *cosine similarity (4).* Aquesta mètrica calcula el cosinus entre els vectors dels termes dels dos elements. (Rahutomo, F., Kitasuka, T., Aritsugi, M. 2012)

$$CosineSimilarity(X, Y) = \frac{X \cdot Y}{|X| \cdot |Y|} \quad (4)$$

3.3.3. Agrupació

Comprobar tots els possibles parells no és una opció. Per tant, prèviament a fer una classificació s'ha de defir un període en el que dos articles poden formar part d'un cas i un valor de similitud per considerar que dos articles són d'un mateix cas.

Per agrupar els articles en casos primer s'han determinat els paràmetres:

- ε similitud mínima per considerar que dos articles formen part del mateix cas.
- T periode en el que es considera que dos articles poden formar part d'un mateix cas

Després de seleccionar aquests paràmetres el següent és construir un conjunt amb totes les parelles possibles i calcula r laseva similitud, per després fer l'agrupació.

Per a seleccionar aquest paràmetres s'han fet proves per diferents valors d'ε comprovant de manera manual parelles de notícies amb diferents valors d'ε i de T.

Crear grups

Per diferenciar els casos dels diferents articles s'han fet servir grafs. Els grafs són estructures formades per un conjunt de vèrtexs amb connexions entre parelles de vèrtexs. El *graph clustering* és una manera d'agrupar vèrtexs tenint en compte l'estructura dels seus enllaços, on hi ha d'haver una gran quantitat d'enllaços a l'interior d'un grup i relativament pocs entre ells. (Schaeffer, S. E. 2007)

En el nostre cas, construïm el graf tal que:

- Cada article és un node.
- Un article s'enllaça amb un altre si la diferencia dels dies de publicació és més petita que T i la seva similitud és més gran que ε.

Aleshores, un cop construït el graf considerem que cada cas és una component connexa[3] d'aquest graf.

3.4. Anàlisis d'articles

Un cop tenim identificats els casos, podem utilitzar els articles. En aquesta part seguirem la mateixa estructura que en l'estudi anterior. Afegint una nova part afegint un nou terme: el temps.

Un dels objectius d'aquesta part és contrastar els resultats trobats anteriorment amb 492 articles, l'altre poder completar aquest anàlisis utilitzant la variable temporal dels casos.

A l'anterior treball es van estudiar els casos considerant tres característiques: el tipus de violència, la relació entre l'agressor i la víctima i el lloc on succeeix. En segon lloc, es va comprovar el tipus d'informació que apareix als articles.

Tot això, comprovant la presència i l'absència de termes i expressions regulars. Els termes i les expressions regulars utilitzades a l'anàlisi es troben a ANNEX 3.

La informació d'aquest apartat s'ha extret de l'estudi anterior, per més informació veure capítol 2.

[3] Subgraf on qualsevol parell de vèrtex estan connectats mitjançant un camí.

3.4.1. Cobertura per casos

Tipus de violència sexual

Els casos de violència sexual, tal i com es van definir a l'anterior estudi, poden ser de tres tipus:

- Assetjament sexual: avenços sexuals no desitjats, sol·licituds de favors sexuals i altres conductes verbals o físiques de naturalesa sexual. L'assetjament no implica penetració.
- Assalt sexual: una violació física psicològica i emocional en forma d'acte sexual, inflingit a algú sense el seu consentiment. Pot implicar el forçament o manipulació d'algú per presenciar o participar en qualsevol acte sexual.
- Abús sexual: un acte de violència infligit per l'agressor contra algú que percep com més febles que ells. És un delicte comès deliberadament amb l'objectiu de controlar i humiliant la víctima.

Un cop comprovada la presència d'expressions regulars en cada article del conjunt de dades, es defineix la violència sexual que més caracteritza cada cas de la següent manera:

- Un cas es considera assetjament si és l'únic tipus de violència sexual i apareix com a mínim en un article del cluster.
- Els articles que contenen termes d'assetjament i assalt es classifiquen com assalt sexual.
- Si un article conté informació relacionada amb abús sexual, el cas es classifica com a abús sexual.

A la següent taula es pot veure aquesta informació de manera resumida.

Taula 3. Assignació tipus de violència sexual

ENTRADA			**SORTIDA**
Assetjament	**Assalt**	**Abús**	**Classificació**
0	0	0	No classificat
0	0	1	Abús
0	1	0	Assalt
0	1	1	Abús
1	0	0	Assetjament
1	0	1	Abús
1	1	0	Assalt

1	1	1	Abús

Relació víctima-agressor

Les categories seleccionades de relacions són:

- Relació: la víctima i l'agressor estaven en una relació afectiva-sexual quan la violència sexual quan la violència sexual es va produir o havia estat en una relació anteriorment.
- Familiar: la víctima i l'agressor tenen un víncle familiar.
- Amic o conegut: la víctima i l'agressor es coneixien abans del crim, considerant, entre d'altres, amica, professors, companys de feina.
- Estrany: quan les persones implicades en la violència sexual eren desconegudes.

Per determinar la relació entre l'agressor i la víctima, s'han considerat el total d'articles que esmenten cada categories.

Les regles utilitzades en aquesta categoría són:

- Si no hi ha cap ocurrència amb cap expressió regular, 'estrany'.
- Si només hi ha una categoria amb un comptador superior a 0, s'assigna aquesta categoria.
- Si hi ha més d'una categoria amb un comprador superior a 0, es asignat el màxim.
- Si no hi ha un màxim únic, és calcula el nombre d'ocurrències només als titulars i es torna a comprovar.

Lloc del crim

L'últim atribut és el lloc on l'acte de violència sexual succeeix. Els diferents tipus són:

- Públic: es van considerar múltiples tipus de llocs públics.
- Lloc de treball.
- Casa: residències.
- Educatiu: lloc on es desenvolupen activitats educatives.
- Oci: llocs d'esbarjo com cinemes, bars, restaurants, discoteques.

Havent calculat el nombre d'ocurrències dels diferents tipus de llocs a cada cluster es considera que el lloc on es va produir el delicte de violència sexual havia de ser el més predominant.

Aleshores el procés de selecció es va fer en tres passos:

1. Regla dicotòmica: Comptar el nombre d'articles en el que es troba cada tipus de lloc i seleccionar el més esmentat. SI hi ha dos llocs o més que apareixen al mateix nombre d'articles passar a la següent etapa.

2. Regla de freqüència de títols i subtítols: seleccionar el tipus de lloc amb la freqüència més alta als títols i subtítols. Si només hi ha un lloc predominant seleccionar; en cas contrari anar al següent pas.

3. Regla de frecuencia corporal: seleccionar el tipus de lloc amb la freqüència més alta al cos dels articles.

Finalment, si cap dels tres passos ha funcional, assignar com ha 'desconegut'.

3.4.2. Anàlisis del contingut

Un com analitzat els casos, també podem analitzar la manera en la què es representa el contingut. Això depèn directament de l'estil d'escriptura del autor.

En aquest anàlisis es pretén obtenir informació sobre la presència o absència d'algun tipus d'informació en el conjunt de dades i la seva ubicació. S'han tingut en compte tots els articles del conjunt de dades.

Els aspectes examinats són: la informació general, els estigmes i les expressions. De la mateixa manera d'abans s'han definit temes i expressions regulars que es troben descrites a l'anex.

Aquest apartat, igual que l'anterior, s'extreu de l'estudi anterior. Per més informació veure capítol 2.

Informació relacionada amb el cas

Aquesta categoria s'utilitza per detectar quin tipus d'informació es presenta amb freqüència als articles sobre violència sexual i si hi ha una associació entre ells.

Les característiques revisades són:

- Actes de violència sexual: si tendeixen a descriure el tipus de violència sexual i les accions realitzades per l'autor.

- Edat: si s'esmenta informació sobre l'edat de les persones esmentades.

- Hora: si s'assenyala la part del dia en la que s'ha produït el delicte.

- Vincle: si és habitual concretar la relació entre les persones implicades.

- Lloc: si els articles proporcionen informació sobre els llocs on s'ha produït el delicte de violència sexual.

Estigmes i expressions

La presencia d'informació estigmatitzada en els articles sobre casos de violència sexual pot influir la manera com la gent pensa sobre les víctimes, els agressors i les situacions. Aquesta secció tendeix a detectar si alguns estigmes generals estan presents a les notícies:

- Origen: si hi ha algun tipus d'informació sobre la l'origen d'una persona de l'article.
- Intoxicat: si l'article conté termes relacionats amb un estat d'embriaguesa i la presència de termes relacionats amb l'alcohol i altres tipus de drogues.
- Roba: termes que impliquen la roba d'algú.
- Vulnerabilitat: si els articles contenen termes que es refereixen a algú en una situació vulnerable.
- Agressor: estigmes sobre l'autor de delictes de violència sexual.

Pel que fa a l'estil d'escriptura:

- Dubte: si sol trobar-se expressions que mostren incredulitat.
- Eufemisme: si els periodistes solen utilitzar expressions suaus o vagues.

4. RESULTATS

4.1. Col·lecció de dades

4.1.1. Obtenció dels tuits

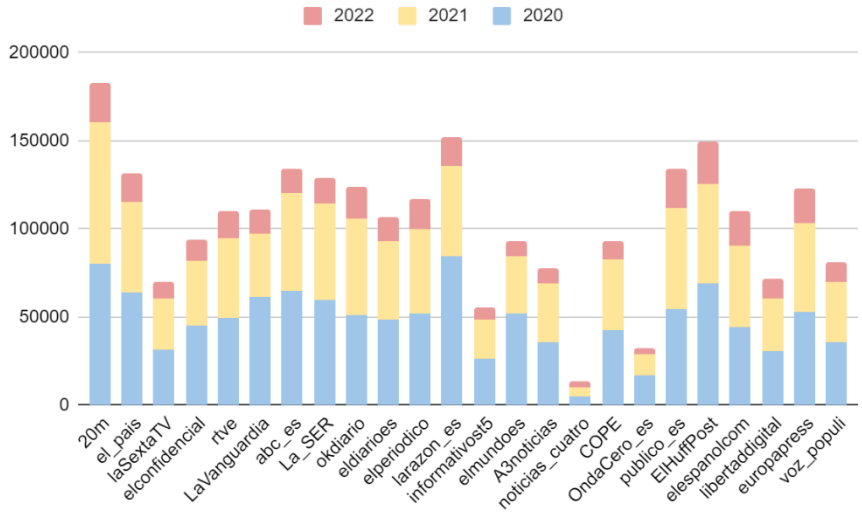
Per aquesta primera fase, crear el conjunt de dades, el primer que fem és obtenir els tuits des del gener 2020 fins el abril 2022 de 24 medis de comunicació espanyols més populars.

Taula 1. Recompte tuits per compte i any

MITJA DE COMUNICACIÓ	COMPTE	2020	2021	2022	TOTAL
20 Minutos	@20m	80286	80290	21971	182547
El País	@el_pais	63745	51169	16357	131271
El Confidencial	@laSextaTV	31209	29539	9074	69822
La Sexta	@elconfidencial	44910	37407	11663	93980
RTVE	@rtve	49010	45820	15465	110295
La Vanguardia	@LaVanguardia	61433	35823	13873	111129
Diario ABC	@abc_es	65202	55323	13490	134015
Cadena SER	@La_SER	59300	55298	14683	129281
Ok Diario	@okdiario	51547	54273	17702	123522
El Diario	@eldiarioes	48510	44317	13473	106300
El Periódico	@elperiodico	52134	47455	17420	117009
La Razón	@larazon_es	84056	51596	16332	151984
Informativos Telecinco	@informativost5	25982	22751	6789	55522
El Mundo	@elmundoes	51568	32755	8920	93243
Antena 3 Noticias	@A3noticias	36032	32626	9140	77798
Noticias Cuatro	@noticias_cuatro	4770	5602	3132	13504
Cope	@COPE	42813	39925	10268	93006
Onda Cero	@OndaCero_es	17216	11776	3245	32237
Diario Público	@publico_es	54753	57058	22449	134260
El Huff Post	@ElHuffPost	68803	56928	23414	149145
El Español	@elespanolcom	44097	46177	19778	110052
Libertad Digital	@libertaddigital	30830	30059	10409	71298
Europa Press	@europapress	52865	50706	19005	122576
Vozpópuli	@voz_populi	35679	34054	11558	81291
		1156750	**1008727**	**329610**	**2495087**

Al definir una franja temporal la quantitat de tuits no està balancejada pels medis però ens aporta informació per poder fer un posterior anàlisis longitudinal i poder fer comparacions amb les estadístiques oficials.

Figura 2 Distribució de tuits per mitjà de comunicació



En la taula anterior podem apreciar la freqüència que tenen els mitjans de comunicació a l'hora de publicar tuits. Per una banda trobem mitjans amb freqüències atípiques, com 20 Minutos o La Razón que tenen freqüències molt elevades, o bé com Noticias Cuatro i Onda Cero que tenen freqüències molt baixes. Encara que la freqüència no és constant pels diferents mitjans es pot veure que dins d'un mateix medi generalment la freqüència és manté al llarg del temps, ja que preserven valors similars al 2020 i 2021.

4.1.2. Classificació dels tuits

Com s'ha explicat anteriorment a l'apartat *3.2.2. Classificació dels tuits* s'ha utilitzat un model de regressió logística supervisada per distingir aquells que són de casos de violència sexual.

Per fer-ho, prèviament s'ha seleccionat un conjunt dels tuits classificats anteriorment per Marilena Budan. S'ha fet servir una tècnica de *subsampling* aleatori per tal d'obtenir al conjunt d'entrenament un 25% de tuits classificats positivament i un 75% de tuits classificats negativament. La primera iteració doncs parteix de 883 tuits positius i 2.649 negatius, un total de 3.532 tuits. Els conjunts d'entrenament i testeig seleccionats han sigut d'un 90% i 10% respectivament.

Amb aquest conjunt ja es pot començar el procés iteratiu definit anteriorment, durant aquest procés s'han anat afegint tuits amb probabilitats pròximes al llindar que millora la puntuació F1, juntament amb falsos negatius i falsos positius.

Finalment obtenim 5.000 mostres, amb 1.090 classificats com a violència sexual i 3.910 que no. Amb la mostra de 500 tweets del conjunt de testeig obtenim els següents resultats pels tweets etiquetats positivament:

Taula 4. Resultats classificador

MÈTRICA	RESULTATS
Precisió	0.917
Reclam	0.869
Exactitud	0.944
AUC	0.971
Millor valor F1	0.893
Millor llindar	0.569

Cada mètrica mesura el funcionament del classificador de diferent manera. Primerament, la precisió mesura el número de mostres classificades correctament, TP/(TP+FP); en canvi el reclam mesura el número de de mostres classificades segons el total, TP/(TP+FN). Si ens fixem en la taula anterior, podem veure que el 91,7% de les mostres s'han classificat correctament i que s'han classificat un 86,9% de les mostres positives.

L'exactitud mesura el percentatge de casos que ha encertat, (TP + TN)/(TP + FP + TN + FN). En aquest cas, podem veure que el nostre model encerta en un 97,7% dels cops.

La següent mètrica que hem utilitzat ha estat el AUC, aquesta és una eina que s'utilitza per mesurar l'encert en els casos binaris es defineix com l'àrea que hi ha sota la corba ROC[4] i mesura la probabilitat de que classifiqui una mostra positiva més alta que una negativa. En el nostre cas, aquesta probabilitat és del 97,1%.

Seguidament, trobem el valor maximitzat de F1. Aquest com hem indicat anteriorment combina les mesures de precisió i reclam en un únic valor. Nosaltres tenim el màxim 0.893 quan el llindar és de 0.569.

4.1.3. Obtenció articles

L'últim pas a seguir per crear el conjunt de dades es extreure els articles des de la web oficial dels mitjans de comunicació dels tuits classificats positivament.

Com s'ha indicat anteriorment no perdem una gran quantitat de tuits positius com passava a l'anterior treball ja que en el nostre cas tenim la URL proporcionada directament per Twitter.

Encara així perdem un 5% dels tuits positius o perquè no redirigeixen a cap URL, perquè ho fan a un article en format vídeo o audio o perquè no ho fan a la seva web principal.

[4] Corba en la què es representa la sensibilitat en funció dels falsos positius per diferents thresholds.

Taula. Recompte notícies per mitjà de comunicació i any

MITJA DE COMUNICACIÓ	COMPTE	2020	2021	2022	TOTAL
20 Minutos	@20m	629	920	306	1855
El País	@el_pais	376	677	250	1303
El Confidencial	@laSextaTV	217	324	141	682
La Sexta	@elconfidencial	156	243	100	499
RTVE	@rtve	56	89	62	207
La Vanguardia	@LaVanguardia	145	325	221	691
Diario ABC	@abc_es	523	537	204	1264
Cadena SER	@La_SER	222	429	173	824
Ok Diario	@okdiario	167	311	110	588
El Diario	@eldiarioes	176	278	118	572
El Periódico	@elperiodico	324	658	383	1365
La Razón	@larazon_es	549	560	154	1263
Informativos Telecinco	@informativost5	361	576	168	1105
El Mundo	@elmundoes	375	395	150	920
Antena 3 Noticias	@A3noticias	252	461	168	881
Noticias Cuatro	@noticias_cuatro	32	6	2	40
Cope	@COPE	155	258	76	489
Onda Cero	@OndaCero_es	28	84	39	151
Diario Público	@publico_es	324	416	115	855
El Huff Post	@ElHuffPost	212	323	143	678
El Español	@elespanolcom	170	253	108	531
Libertad Digital	@libertaddigital	57	107	38	202
Europa Press	@europapress	271	398	170	839
Vozpópuli	@voz_populi	86	72	54	212
		7883	**10721**	**5475**	**18016**

La distribució dels articles en aquest cas no és uniforme ja que la distribució de tuits a la taula x tampoc ho era. Però és proporcional al número de tuits recol·lectats anteriorment. Els medis 20 minutos, El Periódico i La Razón destaquen per la quantitat d'articles que tenen per sobre de la resta.

4.2. Classificació de casos

4.2.1. Obtenció paràmetres

El primer que hem de fer per classificar els casos és definir els paràmetres T i ε. Després de varies proves T = 30 i ε = 0.65.

Per fer-ho s'han provat diferents valors de ε per les possibles T 10, 15, 30 i 60. I s'han comprovat manualment les parelles resultants. Quan s'arriba a T=30, podem observar que encara que augmentem aquest valor el nombre de resultats coincidents no varien massa per les T 30 i 60 i per a qualsevol valor de ε. Per tant (per optimitzar el codi) s'ha escollit T = 30.

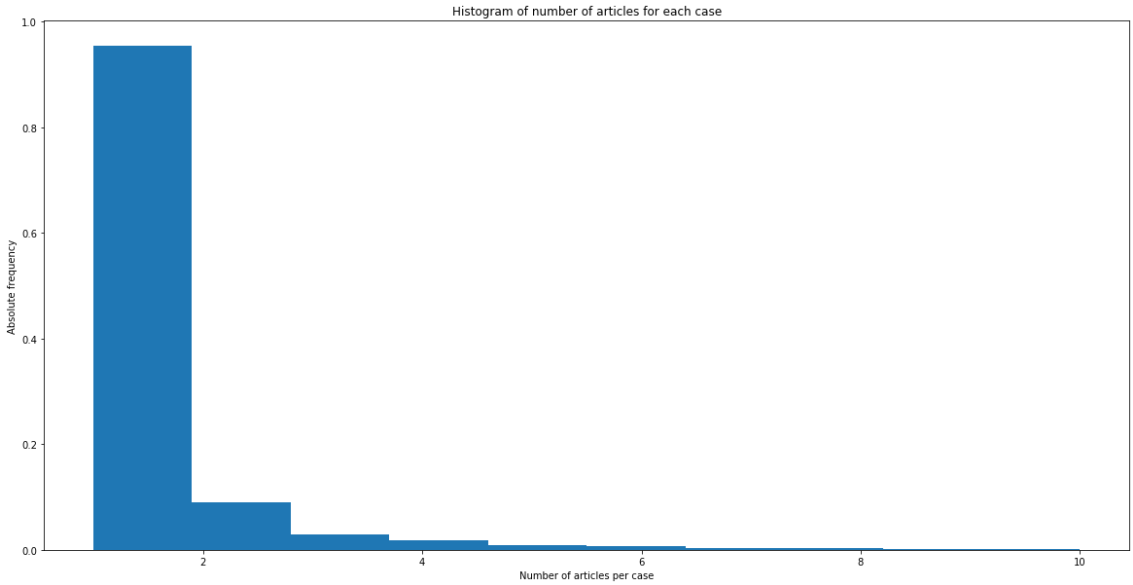
El valor ε inicialment l'havia defint 0,8 ja que a la mostra petita comprovada manualment era el valor que classificava els casos amb més exactitud. Però com estem fent una agrupació a través de grafs conexos, això provocaba que hi hagués a components connexes amb molt de soroll i s'ha reduït al 0,65.

4.2.2. Agrupació de casos

Els articles d'un mateix cas s'agrupen utilitzant grafs com s'ha detallat a l'apartat 3.3.3. Amb els paràmetres seleccionats anteriorment podem agrupar els articles segons el cas que tracten.

Seguint el procediment descrit anteriorment i acotant el nombre de possibles articles que pertanyen a un cas, s'han comprovat 67.564 parelles d'articles. D'aquestes parelles s'han identificat 11.286 casos diferents. La seva distribució és la següent:

Taula 5. Histograma distribució número d'articles per cas

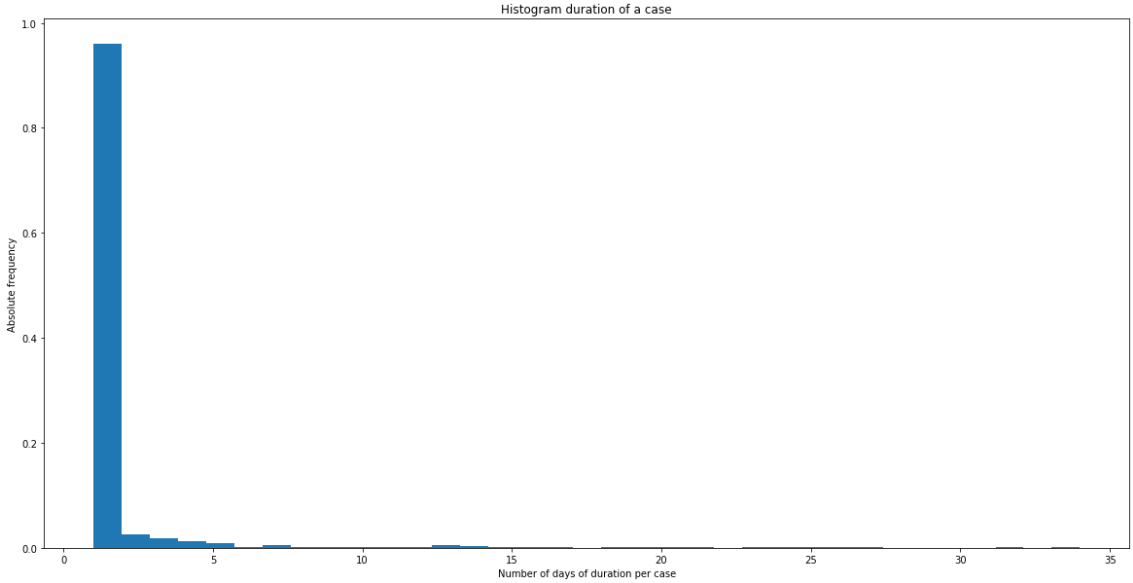


En aquest histograma s'ha limitat els números d'articles a 10 pero la nostre mostra presenta alguns outliers ja que té casos que arriben fins a 28 articles. Aquests casos tenen més soroll degut que utilitzem grafs connexos per classificar-los.

Per exemple, el cas número 415 parla d'un agresor que després de complir la seva condena va tornar a reincidir. Aquest cas té 15 articles relacionats de diferents medis i no s'aprecia gaire soroll.

Per poder veure durant quant de temps es parla d'un cas només tindrem en compte els casos que tenen més d'un article que són 1.625.

Taula 6. Histograma duració d'un cas



En l'anterior histograma podem veure la distribució del número de dies que dura un cas als medis espanyols.

4.3. Anàlisis d'articles

4.3.1. Cobertura per casos

Un cop definides les expressions regulars definides a l'apartat *3.4.1. Cobertura per casos* i les executem. Obtenim els següents resultats:

Tipus de violència sexual

Taula 7 Resultats tipus de casos segons violència sexual

Tipus	Número de casos	Percentatge de casos
Abús	1.118	68,9%
Assalt	392	24,1%
Assetjament	65	4,1%
No classificat	47	2,9%

Relació víctima-agressor

Taula 8 Resultats tipus de casos segons víctima-agressor

Tipus	Número de casos	Percentatge de casos
Conegut	501	30,9%

En relació	149	9,2%
Familiar	232	14,3%
Desconegut	682	43,2%
No classificat	54	3,3%

Lloc del crim

Taula 9. Resultats casos segons lloc de crim

Tipus	Número de casos	Percentatge de casos
Educació	40	2,47%
Casa	185	11,4%
Oci	276	17,0%
Públic	860	53,0%
Treball	32	2,0%
No classificat	80	1,9%

4.3.2. Anàlisis de contingut

Un cop hem analitzat els casos podem passar a executar les expressions regulars de contingut. Aquests han estat els resultats:

Taula 10. Resultats anàlisis de contingut

Tipus	Característiques	Total	Títol	Subtítol	Cos
Actes de violència sexual	Assalt	63,36%	33,62%	20,86%	61,75%
	Assetjament	34,42%	7,16%	6,19%	33,90%
	Abús	59,89%	32,79%	17,00%	57,89%
Informació relacionada amb el cas	Temps	24,67%	0,29%	1,68%	24,57%
	Edat	94,81%	40,35%	38,96%	94,16%
	Treball	21,34%	0,76%	1,26%	21,18%
	Casa	50,05%	2,96%	6,14%	19,70%

	Lloc educatiu	15,39%	1,77%	2,05%	15,03%
	Lloc d'oci	71,44%	5,88%	8,86%	70,59%
	Lloc públic	79,80%	30,98%	19,66%	78,52%
	Relació	40,16%	8,98%	8,42%	39,77%
	Conegut	55,12%	6,41%	8,43%	54,52%
Estigmes	Origen	46,9%	4,67%	6,10%	46,24%
	Intoxicació	16,88%	1,42%	1,93%	16,67%
	Roba	12,90%	0,34%	0,84%	12,75%
	Agressor	4,97%	0,83%	0,38%	4,73%
	Vulnerabilitats	76,7%	28,05%	22,9%	75,64%
Expressió	Eufemismes	14,35%	0,10%	0,70%	14,26%
	Dubte	53,60%	11,56%	10,85%	51,75%

5. DISCUSIÓ I CONCLUSIONS

En aquesta secció l'objectiu és extreure informació i conclusions de les troballes als apartats anteriors durant el desenvolupament del projecte.

5.1. Col·lecció de dades

El primer objectiu i el més important, era ampliar el conjunt de dades anterior. Creant un conjunt de dades amb els articles de caràcter de violència sexual publicats pels mitjans de comunicació espanyols des del gener 2020 fins abril 2022.

Aquesta col·lecció es va iniciar col·lectant 2.495.087 tuits dels 24 mitjans de comunicació generals més importants a espanya.

Un cop recol·lectats tots els tweets el següent pas és classificar-los, per posteriorment obtenir 18.016 articles de les URL que tenen enllaçades.

5.2. Agrupació dels articles

Un cop tenim tota la col·lecció d'articles el següent pas és agrupar-los segons cas, ja que diferents mitjans parlen sovint d'un mateix cas. Podem veure del que més temps es parla són 28 dies, però en la seva majoria només és parla d'ells durant una setmana.

Els casos estan formats per pocs articles excepte en casos puntuals que són molt més mediàtics que la resta i es veuen representats en una gran quantitat de mitjans durant un període de temps més llarg.

5.3. Anàlisis d'articles

Per últim, referent a la anàlisis d'articles, podem veure que en comparació a l'estudi anterior els resultats són molt semblants i no hi ha variàncies significatives.

5.3. Futurs estudis

Seria interessant aprofitant la col·lecció de dades creada fer un anàlisis temporal i resoldre preguntes com: de quins tipus de casos és parla durant més temps?

Per altre banda, seria interessant fer un anàlisis automàtic per tal de diferenciar els casos de forma automàtica en comptes de fer-ho a través d'expressions regulars.

També es podria millorar el l'agrupació per casos, posant un criteri de similitud més baix i utilitzant tècniques per treure el soroll.

Bibliografia

Aizawa, A. 2003. Information processing & management. *An information-theoretic perspective of tf-idf measures*. 45-65. https://doi.org/10.1016/S0306-4573(02)00021-3

Ahmad, A. N. 2010. Is Twitter a useful tool for journalists? Journal of media practice 11(2): 145–155.

Amoedo, A., Vara-Miguel, A., Negredo, S., Moreno, E., Kaufmann, J. (2021). Digital News Report España 2021. Pamplona: Servicio de Publicaciones de la Universidad de Navarra. 85-90. https://doi.org/10.15581/019.003

Budan, M., Castillo C. (2022) The Coverage of Sexual Violence in Spanish News Media. ICWSM Workshop on Data for the Wellbeing of the Most Vulnerable. https://doi.org/10.36190/2022.83

Bernárdez-Rodal, A. López-Priego, N. Padilla-Castillo, G. (2021). Cultura y movilización social contra la violencia sexual a través de Twitter: el caso del fallo judicial "#LaManada" en España. *Revista Latina de Comunicación Social*, 79, 237-262. https://www.doi.org/10.4185/RLCS-2021-1502

González, G., Jiménez, M.L. (2018). Medios de comunicación y violencia sexual : crítica y propuesta de su tratamiento informativo desde un enfoque feminista. En Investigación y género. *Reflexiones desde la investigación para avanzar en igualdad : VII Congreso Universitario Internacional Investigación y Género (311-323), Sevilla: SIEMUS (Seminario Interdisciplinar de Estudios de las Mujeres de la Universidad de Sevilla).*

Harrell, F. E. (2015). Binary Logistic Regression (pp. 259–267). https://doi.org/10.1007/978-3-319-19425-7_10

IAB Spain. (2021). Estudio de Redes Sociales 2021. https://iabspain.es/estudio/estudio-de-redes-sociales-2021/

Instituto Nacional Estadístico. (2021). Estadística de violencia doméstica y violencia de género. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176866&menu=ultiDatos&idp=1254735573206

Martínez, H., Sánchez, G. (2022). La influencia de Twitter en la agenda setting de los mºedios de comunicación. Revista de Ciencias de la Comunicación e Información, 27, 1-21. https://doi.org/10.35742/rcci.2022.27.e136

Newman, N.; Fletcher, R.; Schulz, A.; Andi, S.; and Nielsen, R. K. 2020. Reuters Institute digital news report. Reuters Institute for the Study of Journalism.

O'Hara S. (2012) Monsters, playboys, virgins and whores: Rape myths in the news media's coverage of sexual violence. *Language and Literature*. 21(3), 247-259. https://doi.org/10.1177/0963947012444217

Patil, L. H., Atique, M. (2013) A novel approach for feature selection method TF-IDF in document clustering, 3rd IEEE International Advance Computing Conference (IACC). 858-862. https://doi.org/10.1109/IAdCC.2013.6514339

Rahutomo, F., Kitasuka, T., Aritsugi, M. (2012). Semantic cosine similarity. In The 7th international student conference on advanced science and technology ICAST

Schaeffer, S. E. (2007). Graph clustering. Computer science review. 27-64. https://doi.org/10.1016/j.cosrev.2007.05.001

Susan, S., Kumar, A. (2020). The balancing trick: Optimized sampling of imbalanced datasets - A brief survey of the recent State of the Art. https://doi.org/10.1002/eng2.12298

World Health Organization. (2021). Violencia contra la mujer. *Datos y cifras*. https://www.who.int/es/news-room/fact-sheets/detail/violence-against-women

Zachary C. L., Charles E., Balakrishnan N. (2014) Thresholding Classifiers to Maximise F1 score. https://doi.org/10.48550/arXiv.1402.1892

ANNEX 1. Dades i codi

Les dades i el codi utilitzat per la realització d'aquest treball es poden trobar a:

https://github.com/BegonaLopez0/Dos-anys-de-not-cies-de-viol-ncia-sexual

ANNEX 2. Col·lecció de dades

Les taules a continuació aporten informació sobre els mitjans de comunicació utilitzats i els resultats de la col·lecció de dades.

MITJÀ DE COMUNICACIÓ	*COMPTE*	*DESCRIPCIÓ*
20 Minutos	*@20m*	*Cuenta oficial de 20minutos, el medio social y ciudadano. Información, análisis y contacto personal con los lectores las 24 horas.*
El País	*@el_pais*	*La mejor información en español. Con nuestra mirada puesta en España, Europa y América. Suscríbete: https://bit.ly/39L9ooB*
El Confidencial	*@elconfidencial*	*Somos independientes. ¿Y tú? http://elconfidencial.com/suscribete/ #SoyConfi*
La Sexta	*@laSextaTV*	*Todo el contenido de laSexta en nuestro perfil, en http://laSexta.com y en http://facebook.com/laSexta*
RTVE	*@rtve*	*Canal oficial en Twitter de RTVE, Radio Televisión Española.*
La Vanguardia	*@LaVanguardia*	*Rigor, calidad, compromiso. Suscríbete al periodismo responsable: http://bit.ly/3unbQrY*
Diario ABC	*@abc_es*	*Diario ABC.*
Cadena SER	*@La_SER*	*La fuerza de la voz*
Ok Diario	*@okdiario*	*El sitio de los inconformistas. Dirigido por @eduardoinda*
El Diario	*@eldiarioes*	*Periodismo a pesar de todo.*
El Periódico	*@elperiodico*	*Enteder +*
La Razón	*@larazon_es*	*Información, innovación y emoción.*
Informativos Telecinco	*@informativost5*	*Perfil oficial de Informativos Telecinco.*
El Mundo	*@elmundoes*	*Cuenta oficial de EL MUNDO.*
Antena 3 Noticias	*@A3noticias*	*Toda la información a un click, en http://antena3noticias.com*
Noticias Cuatro	*@noticias_cuatro*	*La redacción de noticias de @cuatro te cuenta la última hora y todas las novedades del día*
Cope	*@COPE*	*Está pasando, estás en COPE 📻 Toda la #información 💻, el mejor equipo de la #radio deportiva 🏆, el mejor #entretenimiento y #podcast 🎙*
Onda Cero	*@OndaCero_es*	*Onda Cero, 24 horas de información y entretenimiento. Te mereces esta radio. Tu #radio*
Diario Público	*@publico_es*	*Twitter oficial de Diario Público.*
El Huff Post	*@ElHuffPost*	*Estamos en: ▶ Instagram: http://bit.ly/2gL30zd ▶ Telegram: https://t.me/elhuffpost ▶ iPhone: http://bit.ly/1AokTa1 ▶ Android: http://bit.ly/1NcE6TE*
El Español	*@elespanolcom*	*Diario digital, plural, libre, indomable, tuyo. Síguenos también en http://tiktok.com/@elespanolcom ¿Eres accionista o suscriptor?*
Libertad Digital	*@libertaddigital*	*Noticias y opinión en la red.*
Europa Press	*@europapress*	*La agencia de noticias privada líder en España*
Vozpópuli	*@voz_populi*	*Periódico digital en español. Información de calidad, grandes exclusivas y el mejor plantel de columnistas. En abierto. vozpopuli@vozpopuli.com*

Comptes de Twitter dels mitjans de comunicació seleccionats

ANNEX 3. Termes i expressions regulars usades per l'anàlisi d'articles

Les taules detallades en aquest annex detallen els termes i les expressions regulars utilitzades per analitzar els articles.

Sexual violence type	Matching terms or expression related with	Regular expression	Matching examples
Sexual assault	Assault	Terms starting with 'agres' or 'agred'	'agresion', 'agresor', 'agredió', 'agredida', 'agrediendo'
	Rape and rapists	Terms containing 'viola' + 'r' / 'cion' / 'dor'	'violar', 'violación', 'violador', 'violadores'
	Penetration	Terms containing 'penetra' + 'r' / 'cion' / 'dor'	'penetrar', 'penetrador, 'penetración', 'penetró'
	Sleeping with someone	Expressions with the following schema 'ocostar'[…]'con'	'acostarse con', ''
	Forcing someone to have sexual practices	Expressions with the following schema 'oblig'[…]'relaciones sexuales'	'obligó a tener relaciones sexuales',
	Prostitution	Terms containing 'prostitu'	'prostituta', 'prostitución', 'prostituirse'
Sexual harassment	Harassment or online harassment	Terms starting with 'acos' or 'ciberacos'	'acoso', 'acosar', 'acosador', 'ciberacoso'
	Grope	'tocamient' or 'manosea'	'tocamientos', 'manosear'
	Intimidation	Terms containing 'indimid' or 'miradas' + adjectives like 'lujuriosas'/'lascivas'/'insistentes	'intimidar', 'intimidante', 'miradas insistentes'
	Extortions	Terms containing 'extors' or 'chantaj'	'extorsionar', chantajear', 'chantajista'
	Nudity	Terms containing 'desnud'	'desnudar', 'desnudó'
Sexual abuse	Abuse	Terms containing 'abus'	'abusar', 'abusador', 'abusó'
	Child abuse	Terms containing 'viol […] menor', expressions such as 'explotación/abuso sexual infantil', and 'relaciones con una menor'	'violar a una menor', 'violador de menores', 'relaciones con una menor', 'abuso sexual infantil'
	Intoxication or disability	Terms containing 'drogad', 'incapaci', 'discapaci'	'drogada', 'incapacidad', 'discapacitada'

Victim-perpetrator bond	Matching terms or expression related with	Regular expression
Relationship	Sexual or affective bond	Terms containing 'matrimonio', 'espos + o/a', 'pareja', 'novi + o/a', 'amante', 'querid +o/a', 'marido', 'su mujer', 'conyugue', 'exnovi +o/a'
Relative	People having a familial bond such as 'son', 'uncle', 'cousin'…	Terms containing 'hij + o/a', 'ti + o/a', 'abuel + o/a', 'sobrin + o/a ', 'ahijad + o/a', 'niet + o/a, 'm/p + adre', 'prim + o/a', 'descendiente', 'herman + o/a'
Acquaintance	People that know each other considering different scenarions such as 'colleague', 'neighbor', 'director', 'professor', 'boss'	Terms containing 'compañer + o/a', 'amig + o/a', 'profesor + a/es', 'alumn + o/a', 'jef + e/a', 'emplead + o/a', 'becari + o/a', 'vecin + o/a', 'maestr + o/a', 'director', 'conocid + o/a', 'entrenador', 'instructor', 'sacerdote'

Type of place	Matching terms or expression related with	Regular expression
Public spaces	Public spaces such as parks, streets, stations, public transport, beaches, gardens, etc.	'avenida', 'parque', 'calle', 'parada', 'bosque', 'plaza', 'carretera', 'puerto', 'estacion', 'jardin', 'fuente', 'montaña', 'espacio publico', 'mirador', 'metro', 'bus', 'tren', 'transporte publico', 'playa'
Workplace	Terms used to refer to workplaces such as offices, shops, coworking spaces, etc.	'oficina', 'trabajo', 'almacen', 'tienda', 'despacho', 'taller', 'coworking', 'gabinete'
House	Ways of referring to homes and types of rooms and spaces of a home	'domicilio', 'casa', 'piso', 'morada', 'hogar', 'vivienda', 'habitacion', 'residencia', 'cocina', 'comedor', 'baño', 'balcon'
Educational places	Place where educational activities occur, such as schools, libraries, high schools, music schools, universities	'universidad', 'escuela', 'biblioteca', 'colegio', 'centro de + educación/enseñanza', 'instituto', 'liceo', 'academia', 'conservatorio', 'guarderia', 'facultad', 'recreo', 'estadio', 'pabellon', 'piscina', 'centro deportivo', 'vestuario'
Leisure spaces	recreation places such as cinemas, theaters, bars, restaurants, shopping centers, nightclubs	'teatro', 'cafeteria', 'discoteca', 'pub', 'bar', 'restaurante', 'centro comercial', 'cine', 'bolera', 'h/m + otel', 'sauna', 'spa', 'piscina', 'gimnasio', 'monumento', 'parque', 'acuario', 'acuarium', 'zoo'

Content analysis	Matching terms or expression related with	Regular expression Matching examples
Time	Expressions referencing parts of the day such as: the same morning, during the afternoon, at midnight, etc.	'la/esta/misma + mañana', ‘'la/esta/misma + tarde’, 'est/al + mediodia', 'atardecer', 'medianoche', 'noche', 'madrugada'
Stigmas	Intoxication: terms referring to an intoxicated state or drugs	'alcohol', 'embriagado', 'droga', 'borracho', 'ebrio', 'bebido', 'alcoholizado', 'fumado', 'estupefacientes', 'intoxicado', 'positivo por/en', 'cocaina', 'consumo de', 'metanfetamina', 'extasis', 'mdma', 'burundanga', 'marihuana', 'porro', 'cannabis', 'hachis', 'sedante', 'speed', 'popper', 'lsd'
	Clothing: terms or expressions referring to the way someone’s clothes	'falda', 'vestido', 'camiseta', 'camisa', 'top', 'tacones', 'leggings', 'pantalones', 'ropa', 'vestid + a/o/as/os + de/con', 'faldilla', 'destapada', 'ceñid', 'escote'
	Origin: terms used for describing someone’s origin or ethnicity such as ‘from the state of’, ‘born in’, ‘latin’, ‘arab’, etc. We append a list with nationalities to the list of regular expressions.	'de origen’, ‘original de', 'su pais + natal/de origen/de procedencia', 'del estado de', 'norte’, ‘sud’, ‘al/el + este’, ‘oeste’, ‘nordeste’. ‘sudeste’, ‘sureste’,’ sudoeste’, ‘suroeste’, ‘noroeste’, 'orient + e/al’, 'occident + e/al', ‘latino', 'hispano', 'arabe', 'mahgrebi', 'caucasico', 'musulman', 'indi + a/o', 'american + o/a', 'europe + o/a', 'asiatic', 'indigena'
	Popular stigmas about the perpetrator: terms demonizing or making the offender look as a sexual predator	'depredador', 'pervertido', 'pervers', 'narcisita', 'solitario', 'enfermo sexual', 'degenerad', 'depravad'.
	Victims’ vulnerabilities or stereotyped situations such as virgins, minors, alone people, etc.	'menor/menores de edad', 'joven', 'indefens', 'fiest', 'desamparad', 'vulnerable', 'abandonad', 'mayor', 'solter', 'promiscu', 'virgen'
Expression	Euphemisms: expressions understating sexual violence acts such as ‘stealing the virginity’, ‘undesired contact’, ‘forcing’, ‘depriving liberty’	'no consentido', 'inapropiad', 'indesead', 'acariciar', 'arrimarse', 'piropear', 'insistir', 'no deseado', 'bajo los efectos de', 'rob[…] + inocencia', 'priv[…] + libertad', 'satisfacer deseos sexuales', 'acceso carnal', 'forz[…] + sex'
	Doubt: Expressions that show a lack of confidence such as ‘presumed’ or ‘alleged’	'supuest[…]/presunt[…] + caso/delito/viola/abus/acos/agre/ victima/responsable/autor/testigo', 'supuestamente', 'acusad + o/a + de', 'presuncion de inocencia', 'acusacion[…] + falsas'