

Aprendentatge automàtic i la violència sexual a Espanya: un enfocament temporal des del 2020 fins al 2022

Gràcia Vallès, Marta

Curs 2022-23

Director: CARLOS ALBERTO ALEJANDRO
CASTILLO OCARANZA

GRAU EN ENGINYERIA MATEMÀTICA EN
CIÈNCIA DE DADES



Universitat
Pompeu Fabra
Barcelona

Escola
d'Enginyeria

Treball de Fi de Grau

Aprendentatge automàtic i la violència sexual a Espanya: un enfocament temporal des del 2020 fins al 2022

TREBALL FI DE GRAU DE
Marta Gràcia Vallès

Director: Carlos Castillo

Grau en Enginyeria Matemàtica en Ciència de Dades

Curs 2022-2023



Universitat
Pompeu Fabra
Barcelona

Escola
d'Enginyeria

Agraïments

Vull donar les gràcies a la meva família i al Guillem pel seu suport emocional durant tota realització d'aquest treball.

A més a més, m'agradaria mencionar especialment al director d'aquest projecte, Carlos Castillo, que ha estat una figura clau en el desenvolupament i l'orientació d'aquest treball. El seu ampli coneixement i experiència han estat fonamentals per encaminar els meus esforços en la direcció correcta, i estic molt agraït pel seu suport constant.

Finalment, vull expressar el meu reconeixement a Marilena Budan i Begoña López, que van fer el projecte del qual ha servit de base per aquest treball.

Resum

Espanya s'ha compromès amb la priorització de l'ampliació de dades i estadístiques sobre la violència sobre la dona per contribuir a la conscienciació social i a l'encert en les actuacions públiques i privades que s'emprenen. Del total de dones de 16 anys o més residents a Espanya, més del 13% han patit violència sexual al menys un cop a la seva vida. Aquest estudi es centra en l'anàlisi dels articles sobre violència sexual publicats pels mitjans de comunicació més populars d'Espanya, dins d'un context en què es vol comprendre com s'aborda i s'informa sobre aquest tema a la societat. L'objectiu principal és desenvolupar un model automatitzat que pugui identificar i agrupar els articles relacionats amb el mateix cas, així com classificar la relació entre l'agressor i la víctima i determinar si és un cas de violència sexual, grupal o individual. Examinant així les característiques principals d'aquests per comparar-ho amb estadístiques oficials del Ministeri d'Igualtat.

Resumen

España se ha comprometido con la priorización de la ampliación de datos y estadísticas sobre la violencia sobre la mujer para contribuir a la concienciación social y al acierto en las actuaciones públicas y privadas que se acometen. Del total de mujeres de 16 o más años residentes en España, más del 13% han sufrido violencia sexual al menos una vez en su vida. Este estudio se centra en el análisis de los artículos sobre violencia sexual publicados por los medios de comunicación más populares de España, dentro de un contexto en el que se quiere comprender cómo se aborda e informa sobre este tema a la sociedad. El objetivo principal es desarrollar un modelo automatizado que pueda identificar y agrupar los artículos relacionados con el mismo caso, así como clasificar la relación entre el agresor y la víctima y determinar si se trata de un caso de violencia sexual, grupal o individual. Examinando así las principales características de estos para compararlo con estadísticas oficiales del Ministerio de Igualdad.

Abstract

Spain is committed to prioritising the expansion of data and statistics on violence against women in order to contribute to raising social awareness and to the success of public and private actions. Of all women aged 16 and over living in Spain, more than 13% have suffered sexual violence at least once in their lives. This study focuses on the analysis of articles on sexual violence published by the most popular media in Spain, within a context in which we want to understand how this issue is addressed and reported to society. The main objective is to develop an automated model that can identify and group articles related to the same case, as well as classify the relationship between the aggressor and the victim and determine whether it is a case of sexual, group or individual violence. Thus, examining the main characteristics of these in order to compare it with official statistics from the Ministry of Equality.

ÍNDEX

1. INTRODUCCIÓ.....	9
1.1 CONTEXT	9
1.2 OBJECTIUS	9
2. ESTAT DE L'ART.....	11
2.1 VIOLÈNCIA SEXUAL A ESPANYA	11
2.2 REPRESENTACIÓ DE LA VIOLÈNCIA SEXUAL ALS MITJANS.....	11
2.3 ANÀLISI AUTOMÀTICA DE NOTÍCIES	12
3. CONJUNT DE DADES.....	15
3.1 PREPROCESSAMENT	16
4. METODOLOGIA.....	17
4.1 PROCEDIMENT	17
4.2 CLASSIFICACIÓ PER CASOS	17
4.2.1 <i>Representació del text</i>	18
4.2.2 <i>Mètriques de similitud</i>	19
4.2.3 <i>Característiques</i>	20
4.2.4 <i>Càlcul de la probabilitat</i>	21
4.2.5 <i>Agrupació</i>	21
4.3 CLASSIFICACIÓ PER TIPUS DE RELACIÓ I NÚMERO D'AGRESSORS	22
4.3.1 <i>Etiquetar la data</i>	22
4.3.2 <i>Generar data sintètica</i>	23
4.3.3 <i>Oversampling i Undersampling</i>	23
4.3.4 <i>Model</i>	24
4.4 ANÀLISI DEL VOCABULARI	25
5. RESULTATS.....	27
5.1 CLASSIFICACIÓ DE CASOS	27
5.1.1 <i>Model</i>	27
5.1.2 <i>Agrupació per casos</i>	28
5.2 CLASSIFICACIÓ PER TIPUS DE RELACIÓ I NÚMERO D'AGRESSORS	29
5.2.1 <i>Relació Víctima – Agressor</i>	29
5.2.2 <i>Número d'agressors</i>	32
5.3 ANÀLISIS DEL VOCABULARI	34
6. DISCUSSIÓ I CONCLUSIONS	37
6.1 DISCUSSIÓ	37
6.1.1 <i>Classificació per casos</i>	37
6.1.2 <i>Anàlisi del vocabulari</i>	39
6.2 CONCLUSIONS I FUTURS ESTUDIS	40
7. BIBLIOGRAFIA	43
ANNEX 1. DADES I CODI.....	47
ANNEX 2. RESULTATS CLASSIFICADORS DE TIPUS DE RELACIÓ AGRESSOR-VÍCTIMA I NÚMERO D'AGRESSORS INVOLUCRATS EN EL CAS	48
ANNEX 3. RESULTATS ANÀLISI DEL VOCABULARI	50

1. INTRODUCCIÓ

1.1 Context

A Espanya, hi ha hagut un augment significatiu dels delictes de violència sexual en l'última dècada. Segons l'Institut Nacional d'Estadística, les dades indiquen que 32.644 dones van ser víctimes de violència de gènere el 2022 xifra que ha augmentat un 8,3% respecte al 2021 on el nombre de víctimes van ser 30.141 (Instituto Nacional Estadístico, 2023). És important tenir en compte que els mitjans de comunicació seleccionen els esdeveniments que es publiquen, la qual cosa pot portar a una representació esbiaixada dels fets reals. Atès que els articles influeixen en l'opinió pública, la manera en com es plasma la violència sexual en els mitjans és un tema rellevant per a la societat ja que, com sabem, la cobertura informativa d'aquestes notícies té el potencial per influir en els coneixements, les creences i les actituds de les persones en relació a la violència sexual (Sreedharan, C., et al, 2021).

1.2 Objectius

Els objectius principals d'aquest treball són els següents:

1. Desenvolupar un model automatitzat capaç d'identificar i agrupar els articles que tracten sobre el mateix cas per veure en quina proporció s'aborden els casos al mitjans.
2. Crear un classificador multiclasse basat en models de processament del llenguatge natural per determinar la relació existent entre l'assaltant i la víctima.
3. Crear un classificador que permeti detectar i determinar si hi ha un o més agressors implicats en un cas específic.
4. Analitzar i comparar el llenguatge emprat en cadascun d'aquests dos casos.
5. Comparar amb les estadístiques nacionals oficials si els mitjans representen amb la mateixa mesura tots els tipus de cas i de quina manera ho fan.

En resum, la finalitat és abordar aquests objectius i proporcionar una solució per a la identificació, classificació i anàlisi de casos.

2. ESTAT DE L'ART

2.1 Violència sexual a Espanya

La violència sexual és una forma de violència contra la dona molt present en la societat. Una de cada cinc dones residents a Espanya de 16 anys o més han patit violència sexual o física per part de parelles, exparelles o tercera (Macroencuesta de Violencia contra la Mujer 2019).

L'estudi estadístic més important a Espanya és la *Macroencuesta de Violencia contra la Mujer 2019*. Aquesta enquesta va recollir per segona vegada informació sobre la violència sexual fora de la parella, permetent així conèixer millor l'extensió real de la violència sexual, ja que, les macroenquestes anteriors no van incloure la mesura d'aquesta fora de la parella, exceptuant la del 2015 (Ministerio de Igualdad, 2018).

Els resultats d'aquest estudi proporcionen informació sobre el tipus de relació entre la víctima i l'agressor (representada a la *Taula 1*) junt amb la freqüència d'agressions individuals i en grup mostrada a la *Taula 2*.

Taula 1: Freqüència dels diferents tipus de vincles entre la víctima i l'autor de la violència sexual

Relació Víctima Aggressor	Freqüència en la Macroencuesta 2019
Desconeugut	10.86%
Conegut	14.03%
Familiar	6.00%
(Ex)Parella	72.20%

Taula 2: Freqüència d'agressions grupals i d'un sol autor involucrat en el cas

Número d'autors involucrat	Macroencuesta 2019
Individual	87.96%
Grupal	12.04%

2.2 Representació de la violència sexual als mitjans

És sabut que la violència sexual té conseqüències molt greus sobre la salut física i psíquica de les víctimes. L'Organització Mundial de la Salut considera la violència sexual com un dels problemes més importants en salut pública i drets humans a escala global (World Health Organization, 2013).

Els mitjans de comunicació exerceixen un paper vital en la formació de l'opinió pública. Tenint en compte que aquests tenen la capacitat de manipular i influir en les idees i creences dels lectors (Fitzpatrick, 2018; Morgan, 2018), és crucial la representació de la violència sexual en aquests per entendre la propagació de prejudicis, mites i estereotips entorn d'aquest tema, així com el llenguatge emprat per a la descripció d'aquest tipus de cas. A continuació presentem uns quants estereotips i mites de l'actualitat.

- Les persones que cometen delictes de violència sexual pateixen problemes de salut mental, per tant, aquestes conductes no han d'estar generalitzades (Walton, 2020).
- Els delictes de violència sexual són duts a terme per gent considerada com "boja" (Walton, 2020).
- En els casos d'abús infantil, les notícies se centren en la "naturalesa depredadora dels agressors" i soLEN descriure com delictes violents comesos per desconeGUTS.
- És més probable que una dona sigui violada per algú desconeGUT que per un coneGUT (Ministerio de Igualdad, 2018).
- Un dels motius per la qual alguns homes agredeixen sexualment a les dones és pel consum d'alcohol i/o drogues (SEXVIOL, 2022).
- Violació real: acte sexual forçat que té lloc en un espai aïllat a altes hores de la nit per part d'algú desconeGUT, i on existeix un alt grau de violència que provoca conseqüències físiques i sexuals en les víctimes (Estrich, 1987).
- Les agressions sexuals en grup són freqüents (SEXVIOL, 2022).

2.3 Anàlisi automàtica de notícies

L'enfocament principal utilitzat per a l'anàlisi automatitzada de notícies és a través del Processament del Llenguatge Natural (NLP, per les seves sigles en anglès). El NLP és una branca de la intel·ligència artificial que proporciona als ordinadors la capacitat d'entendre textos i paraules de la mateixa manera que un ésser humà (Torfi, A., et al., 2021). El NLP combina la lingüística computacional amb models estadístics, aprenentatge automàtic i aprenentatge profund per aplicar-se a una alta gamma d'àrees com la traducció automàtica, l'etiquetatge gramatical, la classificació de textos, l'extracció d'informació o l'anàlisi de sentiments (Li, Q., et al., 2022).

Una de les millors llibreries per treballar NLP en Python és NLTK (*Natural Language Toolkit*), aquesta inclou altres llibreries per realitzar subtasques com per exemple la tokenització, anàlisis d'oracions, segmentació de paraules, lematització, entre d'altres. La classificació de text requereix preprocessar el text d'entrada per a entrenar els models tradicionals. Aquesta representació del text, pretén expressar el text preprocessat d'una forma que sigui molt més fàcil pels ordinadors i minimitzi la pèrdua d'informació, com *Bag-Of-Words* (BOW) (Zhang, Y., et al., 2010), *N-gram* (William B. C., et al., 1994), *Term Frequency-Inverse Document Frequency* (TF-IDF) (Baeza-Yates, R., Ribeiro-Neto, B., 1999) i *word2vec* (Mikolov, T., et al., 2013).

Una vegada hem obtingut la representació vectorial de la col·lecció de dades, es poden aplicar diferents tipus d'anàlisis en funció de l'estudi. Encara que no s'hagi dissenyat específicament per a tasques de classificació de text, la representació codificada bidireccional a partir de transformadors (BERT) (Devlin et al., NAACL 2019), s'ha utilitzat en models de classificació de texts donada la seva eficàcia en nombrosos conjunts de dades.

BERT (sigles en anglès de *Bidirectional Encoder Representations from Transformers*), és un algoritme d'aprenentatge profund, desenvolupat per l'equip de *Google AI Language*, que representa paraules en un *embadded space*, centrant-se en la comprensió del llenguatge capturant així coneixements sintàctics i semàntics de les paraules (Rogers et al., 2020). Històricament, els models de llenguatge només podien llegir l'entrada del text de manera seqüencial, fos d'esquerra a dreta o de dreta a esquerra, però no podien fer les dues coses al mateix temps. BERT, en canvi, està dissenyat per llegir en ambdues direccions alhora. Aquesta capacitat, habilitada per la introducció de Transformers, es coneix com a bidireccionalitat (Devlin et al., NAACL 2019).

3. CONJUNT DE DADES

En aquest treball, s'ha utilitzat el conjunt de dades generat per Begoña López en el seu treball final de grau¹, basat en el Digital News Report Espanya 2021 (Amoedo, A., 2021). Els articles es troben organitzats en carpetes segons l'any i el mes de publicació. A més, els articles estan en format NewsML-G2, una forma estandarditzada de desar-los en format XML, establerta per l'International Press Telecommunications Council² (IPTC) que proporciona estàndards oberts per als mitjans de comunicació.

Taula 3: 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] ></infoSource>
    </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>
```

¹ <https://github.com/BegonaLopez0/Dos-anyys-de-noticies-de-violencia-sexual>

² <https://iptc.org>

Com es pot observar a la *Taula 3*, les dades emmagatzemades inclouen diversos camps que es poden classificar en les següents categories:

- Identificador de l'article: compost a partir de "urn:[MEDIA]:[EXTRACTION DATE]:[TWEET ID]"
- Extracció i dades d'origen: especificant la data de l'extracció i l'ID del tuit on es va obtenir l'URL de l'article.
- Informació de l'article: font de l'URL, data i hora de publicació, la font que va publicar l'article, el país en què s'ha publicat i l'autor de l'article.
- Títol de l'article: el titular i el subtítol.
- Text de l'article: el cos de l'article.

3.1 Preprocessament

El següent pas consisteix a processar tots els articles XML corresponents a cada mes i any, amb la finalitat d'aconseguir un dataframe amb tots aquests. Una vegada extret el text, procedirem a normalitzar el títol, el mitjà, l'encapçalat i el contingut de l'article, posant en minúscula tots els caràcters, eliminant signes de puntuació i stopwords³, de manera que el dataframe final tingui l'estructura representada en la *Taula 4*.

Taula 4: Estructura del dataframe dels articles normalitzats

tweet_id	media	title	headline	url	article	publication_date
1212203735650 328577	20m	britanica detenida falso testimonio...	joven britanica 19 anos mes julio detenida	https://www.20minutos.es/noticia/...	joven britanica 19 anos mes julio detenida...	2019-12-31
1212467642428 280832	20m	parto nina 13 anos violada padre	nina brasileña 13 años fallecido...	https://www.20minutos.es/noticia/...	nina brasileña 13 años fallecio comienz...	2020-01-01
1212826760234 127360	20m	3 jovenes acusados violar 3 hermanas estadouni...	detenido jueves tres jóvenes acusados violaci...	https://www.20minutos.es/noticia/...	detenido jueves tres jóvenes acusados violaci...	2020-01-02
1213232973715 267584	20m	claves juicio abusos sexuales despertaron	juicio supuestos abusos sexuales todopoderoso ...	https://www.20minutos.es/noticia/...	juicio supuestos abusos sexuales todopoderoso ...	2020-01-03
1213348259076 415488	20m	investiga escritor frances ...	abierto viernes investigacion escritor frances...	https://www.20minutos.es/noticia/...	abierto viernes investigacion escritor frances...	2020-01-03

³ Paraules utilitzades amb freqüència que no proporcionen informació útil per al propòsit computacional, com per exemple, «i», «o», «el», entre d'altres.

4. METODOLOGIA

4.1 Procediment

La metodologia seguida per aconseguir els objectius establerts a la secció [1.2 Objectius](#) es resumeix en els següents apartats:

- a) Classificació per casos
- b) Classificació per tipus de relació i número d'agressors
- c) Anàlisi del vocabulari

4.2 Classificació per casos

Atès que els casos de violència sexual tenen una àmplia cobertura, és important entendre que dins de tots els articles obtinguts, un mateix cas pot estar representat en diversos articles del mateix o diferent mitjà de comunicació.

Per tal d'agrupar tots els articles que tractin sobre el mateix cas, tindrem en compte dues consideracions. La primera és que el nostre conjunt de dades, en tractar tots els articles sobre el mateix tema, conté un vocabulari similar i no és una tasca trivial agrupar els articles que presenten informació sobre el mateix cas. I finalment és que el volum de dades que tenim és molt gran i, per tant, hem d'intentar fer el codi de la manera més eficaç i òptima possible.

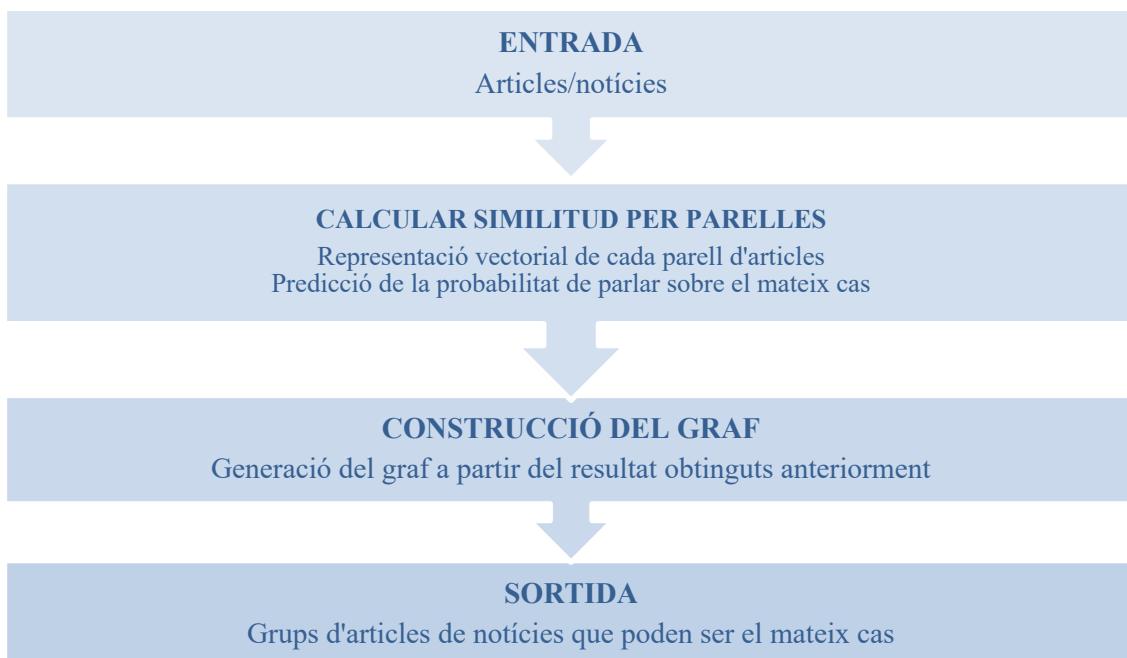


Figura 1: Procés per agrupar articles de notícies que són el mateix cas

Aquest procés implica diferents passos: representar cada parell d'articles com a vectors de característiques, predir la probabilitat de ser el mateix cas d'una manera supervisada, i agrupar tots els articles en un graf mitjançant la probabilitat predita prèviament.

La metodologia emprada en aquest apartat és diferent de la de l'anterior estudi per tal de guanyar més precisió a l'hora d'agrupar els diferents casos.

4.2.1 Representació del text

Hi ha múltiples maneres de representar documents de text, des de la representació de la bossa de paraules fins a la incrustació de documents (Sriram, 2020). Per comparar parells de documents, cada article ha estat representat de diverses maneres per aprofitar la informació que cada representació proporciona.

a) TF-IDF

El *Term Frequency Inverse Document Frequency* és una mesura utilitzada en la recuperació d'informació i la mineria de text per a avaluar la importància de les paraules en una col·lecció de documents. Consta de dues parts: TF (freqüència del terme) i IDF (freqüència inversa del document).

El TF mesura quantes vegades apareix una paraula específica en un document. Les paraules amb un alt valor de TF són considerades importants en aquest document. Per altra banda, el DF compta quantes vegades apareix una paraula en tota la col·lecció de documents i, per tant, les paraules amb un alt valor de DF no són considerades rellevants, ja que apareixen comunament en tots els documents.

El IDF és l'invers de DF i s'usa per a mesurar la importància de les paraules en tots els documents. Així doncs, un alt valor de IDF té una gran importància perquè són menys comuns en tota la col·lecció (Kim, SW., Gil, JM 2019).

Donat un terme t i un document d el TF-IDF s'obté amb l'expressió (1), on N és el nombre total de documents.

$$TF - IDF_{t,d} = tf_{t,d} \cdot idf_t = tf_{t,d} \cdot \log\left(\frac{N}{df_t}\right) \quad (1)$$

- $tf_{t,d}$ 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

El valor de TF-IDF augmenta quan una paraula té una alta freqüència en un document específic i és poc comú en el conjunt total de documents. Això ens ajuda a identificar les paraules clau més rellevants en cada document.

b) Vectors de paraules

El *Word embedding* és un enfocament amb el qual representem documents i paraules en un espai vectorial continu (Almeida, F., Geraldo, X., 2023). En aquest espai, les paraules semànticament o sintàcticament relacionades han d'estar situades a la mateixa àrea. Un avantatge important de la incrustació de paraules és que la seva formació no requereix un corpus etiquetat (A. G. D'Sa, et al., 2020).

- Fast Text

El *FastText* es basa en el model *skip-gram*⁴, on cada paraula es representa com una bossa de n-grams (A. Joulin, et al., 2016). S'associa una representació vectorial a cada n-gram; les paraules es representen com la suma d'aquestes representacions. El *FastText* és capaç de proporcionar un *embedding* per a paraules mal escrites, paraules rares o paraules que no estaven presents en el conjunt d'entrenament, perquè *FastText* utilitza la tokenització de paraules amb n-grams de caràcters. (A. G. D'Sa, et al., 2020).

- BERT (BETO)

BERT, prèviament explicat *l'apartat 2.3*, és multilingüe i per tant el podríem utilitzar però, en aquest projecte s'ha fet servir el model preentrenat només amb text en castellà, ja que a diferència del multilingüe, amb aquest s'obtenen millors resultats (Cañete et al., 2020).

4.2.2 Mètriques de similitud

- Jaccard coeficient

El coeficient de similitud de Jaccard entre dos conjunts de dades és el resultat de la divisió entre el número de característiques que són comunes a tots dos textos dividit entre el número de propietats, com podem veure a continuació (Niwattanakul, S., et al., 2013).

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (2)$$

S'ha de tenir en compte que aquesta mètrica utilitza un enfocament de *bag-of-words*⁵, ja que no considera la freqüència dels atributs.

⁴ *Skip-gram* és un mètode per aprendre word embeddings, que són representacions contínues, denses i de dimensions petites de paraules en un vocabulari.

⁵ *Bag-of-Words* és una tècnica que representa documents de text com a conjunts de paraules desordenades, sense considerar l'ordre o la estructura gramatical.

- Cosine Similarity

Donada una representació vectorial de dos elements, la seva similitud de cosinus és el cosinus de l'angle entre ambdós vectors. La mètrica es pot calcular a partir dels vectors com el seu producte de punts normalitzat pel mòdul del seu producte creuat (Huang, 2008).

$$\text{CosineSimilarity}(\vec{X}, \vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X} \times \vec{Y}|} \quad (3)$$

- Word Movers

Word Movers Distance proporciona una funció de distància efectiva utilitzant *word embeddings* preentrenats (Sato, R., et al., 2022). Donats els *word embeddings* de dos textos diferents, la distància del *word movers* calcula la distància entre els textos com la distància acumulada mínima que els *embeddings* d'un text han de viatjar per transformar-se en la representació del segon text. Com que es calcula fent servir *word embeddings*, la mètrica considera els canvis que s'han de fer a un nivell semàntic i sintàctic (Bahrdwaj, Saksham; Saxena, 2019).

4.2.3 Característiques

La classificació per casos s'ha realitzat de forma aparellada, la qual cosa significa que la matriu de característiques conté un vector per a cada parell d'articles, i s'ha utilitzat per classificar de manera binària si aquests dos articles presenten informació sobre el mateix cas. Les característiques han estat creades combinant tècniques de representació de text amb mètriques de similitud, presentades a les seccions 3.3.1 *Representació del text* i 3.3.2 *Mètriques de similitud*.

- *Word Movers Distance* del *embedding* de *FastText* del títol
- Coeficient de Jaccard del títol i del titular
- *Cosine similarity* de TF-IDF considerant totes les parts de l'article
- *Cosine similarity* del *BETO embedding* considerant totes les parts de l'article

Considerant les propietats dels articles de notícies en lloc del seu contingut com hem fet fins ara, fem servir la diferència absoluta de dies entre la publicació d'ambdós articles afegint així nova informació a la matriu de característiques. Amb aquesta característica definim un període de màxim tres dies perquè dos articles puguin ser d'un mateix cas.

4.2.4 Càlcul de la probabilitat

El model de classificació utilitzat per aquesta tasca és una regressió logística supervisada entrenada amb les dades etiquetades del treball anterior de Marilena Budan (Budan, M., Castillo, C. 2022). L'objectiu d'aquesta secció és predir la probabilitat de pertànyer a la classe 1 (tots dos articles són sobre el mateix cas) amb les característiques descrites anteriorment a l'apartat 4.2.3 *Características*. El model *Binary logistic regression* calcula aquesta probabilitat a través de funció *sigmoid* (4).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

La sortida d'aquesta funció és la probabilitat de pertànyer a la classe 1. Per tal de determinar la classe final, s'aplica un *threshold* a aquest valor de probabilitat, normalment 0,5. Les probabilitats més grans que aquest threshold pertanyen a la classe 1 i les probabilitats més baixes, a la classe 0 (Kleinbaum, D.G., Klein, M., 2010).

4.2.5 Agrupació

Per tal d'agrupar els diferents articles en casos s'ha utilitzat 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 seva similitud és més gran que 0,85.

Després de construir el graf, considerem que cada component connexa⁶ d'aquest graf representa un cas diferent.

⁶ En teoria de grafs, un component connex és un subgraf en què qualsevol parell de vèrtexs estan connectats mitjançant un camí.

4.3 Classificació per tipus de relació i número d'agressors

Aquesta secció es compon de dues parts. En primer lloc, es troba un classificador multiclassificació per a identificar la relació entre la víctima i l'agressor. En segon lloc, un classificador que determina el tipus d'agressió sexual, ja sigui grupal, individual, etc.

Per a abordar aquests dos casos, hem seguit la mateixa estructura descrita en la *Figura 2*.



Figura 2: Procés de classificació

4.3.1 Etiquetar la data

Per entrenar el model hem etiquetat 900 articles manualment amb la seva etiqueta associada. La codificació utilitzada en cada cas per a l'assignació d'etiquetes es mostra a les següents taules.

Taula 5: Criteri utilitzat per etiquetar els articles en funció de la relació víctima-assaltant.

VALOR	RELACIÓ VICTIMA - ASSALTANT
0	Desconegut
1	Parella
2	Familiar
3	Conegut

Taula 6: Criteri utilitzat per etiquetar els articles en funció del número d'agressors.

VALOR	NÚMERO D'AGRESSORS
0	Individual
1	Grupal (més d'un agressor)
2	Altres ⁷

Després de l'etiquetatge manual aconseguim la següent quantitat d'articles per cada etiqueta.

⁷ Articles que no s'especifica el número d'agressors o articles informatius (sobre lleis per exemple)

Taula 7: Número d'articles per cada etiqueta en relació víctima assaltant.

VALOR	NÚMERO D'ARTICLES
0 - Desconegut	478
1 - Parella	115
2 - Familiar	173
3 - Conegut	118

Taula 8: Número d'articles per cada etiqueta en número d'agressors.

VALOR	NÚMERO D'ARTICLES
0 - Individual	539
1 - Grupal	219
2 - Altres	126

Com podem observar en ambdues taules – *Taula 7 i 8* – les dades es troben bastant *unbalanced*⁸, per tant, procedirem a aplicar diferents tècniques de *resampling*⁹.

4.3.2 Generar data sintètica

Després del procés d'etiquetatge, separam les dades en dos conjunts, un primer conjunt d'entrenament (80%) i un segon conjunt de testatge (20%). Aquest primer grup d'entrenament l'utilitzarem per a generar nova data sintètica a través del model CTGAN (Conditional Tabular Generative Adversarial Network) (Lei Xu et al., 2019), aconseguint així augmentar la data d'entrenament a 1.414 articles. Tot i això, encara tenim les nostres dades molt desequilibrades en ambdós casos i, per tant, aplicarem les tècniques d'*undersampling* i *oversampling* explicades en el següent apartat 4.3.3.

4.3.3 Oversampling i Undersampling

L'*undersampling* i l'*oversampling* són tècniques usades per a abordar el desequilibri de dades en problemes d'aprenentatge automàtic. Per una banda, l'*undersampling* és un mètode que busca equilibrar la distribució de classes mitjançant l'eliminació aleatòria d'exemples de la classe majoritària. L'objectiu és tractar d'equilibrar el conjunt de dades,

⁸ Un conjunt de dades *unbalanced* és un conjunt on el nombre d'observacions pertanyents a una classe és significativament major o menor que les pertanyents a les altres classes.

⁹ El *resampling* és un mètode que implica replicar mostres del conjunt de dades d'entrenament.

no obstant això, un desavantatge important és que pot descartar dades que podrien ser importants. D'altra banda, l'*oversampling* és un mètode que busca equilibrar la distribució de classes mitjançant la replicació aleatòria d'exemples de la classe minoritària. Tanmateix, aquest mètode pot augmentar el risc d'*overfitting*¹⁰, ja que crea còpies exactes dels exemples de la classe minoritària i això pot portar al fet que el model generi regles d'aprenentatge per a aquestes (Kotsiantis et al., 2006).

D'aquesta manera aconseguim els següents resultats:

Taula 9: Número d'articles per cada etiqueta en relació víctima assaltant post resampling.

VALOR	NÚMERO D'ARTICLES
0 - Desconegut	500
1 - Parella	250
2 - Familiar	350
3 - Conegut	250

Taula 10: Número d'articles per cada etiqueta en número d'agressors post resampling.

VALOR	NÚMERO D'ARTICLES
0 - Individual	600
1 - Grupal	400
2 - Altres	400

4.3.4 Model

Abans de crear i entrenar el model, hem de transformar el text de tal manera que el model sigui capaç de processar-lo. D'aquesta manera el que fem és *tokenitzar*¹¹ tot el text de l'article (títol, titular i cos de l'article). Limitem els *tokens* a 512, ja que són el nombre màxim de *tokens* que el model BERT pot processar, tot i això, en els articles, la informació més rellevant sol estar al títol, al titular o a l'inici del cos de l'article, per tant, aquesta limitació no ens hauria de perjudicar.

¹⁰ L'*overfitting* ocorre quan un model d'aprenentatge automàtic s'ajusta massa a les dades d'entrenament i perd capacitat de generalització.

¹¹ La *tokenització* divideix el text sense format (per exemple, una oració o un document) en una seqüència de tokens, com a paraules o subparaules.

A la següent taula podem veure una representació del model d'aprenentatge automàtic utilitzat en ambdós casos on, la diferència entre ells és la dimensió de la sortida. En el model de classificació per a determinar la relació entre la víctima i l'agressor (*Taula 11*), la sortida és de 4 dimensions, en canvi, en el model que determina el número d'agressors, la sortida és de 3 dimensions.

Taula 11:Model d'aprenentatge automàtic.

Input Layers [(None, 250)]	Total params: 109,853,956
TFBertModel	Trainable params: 109,853,956
Dense (None, 4)	

4.4 Anàlisi del vocabulari

En el camp de recerca del *Natural Language Processing* (NLP) hi ha diverses tasques que contribueixen a entendre el text. Aquestes tasques poden manipular el llenguatge, com per exemple el procés de tokenització, i per tant es poden utilitzar en altres implementacions, per tal d'extreure informació sintàctica o semàntica. Una d'aquestes tasques per als components sintàctics és *Part of Speech Tagging* (POS Tagging)¹². El *POS Tagging*, s'ha desenvolupat fent servir la llibreria spaCy. SpaCy és una llibreria de *Natural Language Processing* de codi obert, que suporta una gran varietat de tasques com per exemple el POS Tagging, el *Named Entity Recognition*, *Dependency Parsing*, etc (Partalidou, E., et al., 2019).

En aquest apartat, el nostre objectiu és analitzar el vocabulari emprat per a descriure els articles i determinar si existeix alguna tendència en la forma en la qual es descriuen els diferents tipus de cas. També, volem examinar quins aspectes es destaquen més quan es tracta d'un cas d'una agressió grupal, una agressió d'un desconegut o d'una agressió d'un familiar, entre altres escenaris.

¹² *Part of Speech Tagging* és un procés on una paraula s'assigna amb una etiqueta del terme gramatical, donat el context en el qual aquesta apareix.

Per tal d'abordar això i ser els més precisos possibles, en analitzar el tipus de vocabulari, primer de tot el classifiquem en funció de la seva categoria gramatical utilitzant el *POS Tagging*, explicat anteriorment, i partir d'aquí, amb cada categoria visualitzem la distribució de paraules amb *Shifterator*¹³ (Ryan J. Gallagher, 2020).

Distingim 4 tipus de categories; adjectiu ('ADJ'), adverbi ('ADV'), nom ('NOUN') i verb ('VERB'), i per cadascuna d'elles, s'ha examinat el vocabulari emprat els següents casos:

- Tipus d'agressió: Individual o grupal
- Tipus de relació entre la víctima i l'agressor
 - (Ex)Parella o desconegut
 - Familiar o desconegut

¹³ Shifterator és un paquet de Python per visualitzar comparacions entre textos a través de *word shifts*, un mètode per extreure quines paraules difereixen entre dos textos i de quina manera ho fan.

5. RESULTATS

5.1 Classificació de casos

5.1.1 Model

El classificador utilitzat per distingir els parells d'articles que proporcionen informació sobre el mateix cas s'ha entrenat amb dades on la proporció de cada etiqueta és notablement diferencial, per tant, per tal d'analitzar el rendiment del model utilitzarem la següent *Taula 12*.

Taula 12: Rendiment del model de classificació per a detectar parells d'articles sobre el mateix cas

MÈTRIQUES	Etiqueta 0 (diferent cas)	Etiqueta 1 (mateix cas)
Precision	0.99	1.00
Recall	1.00	0.65
F1-score	1.00	0.78
Accuracy		0.99
Support	1784	31

És important tenir en compte que l'*accuracy*¹⁴ pot no ser la mètrica més adequada quan les classes estan desequilibrades. El model de regressió logística mostra un rendiment excel·lent per a la classe 0, en canvi, per la classe 1 podem veure una *precision*¹⁵ del 100%, el que implica que totes les instàncies classificades com classe 1 son realment d'aquella classe. No obstant això, la *recall*¹⁶ és del 65%, indicant així que el model només identifica el 65% de les instàncies reals de la classe 1. Cal tenir en compte que el suport de la classe 1 és baix i que, per tant, tenim una quantitat petita de mostres d'aquesta classe.

La *Figura 3* mostra la importància de cada característica deduïda dels pesos que assigna el model. Les característiques més rellevants observades són TF-IDF i el coeficient de Jaccard. A més a més, podem veure com les característiques que avaluen la distància entre dos documents (Word Movers i BETO), han estat assignades amb pesos negatius.

¹⁴ La fracció de prediccions que el model realitza correctament.

¹⁵ El número de prediccions correctes de totes les prediccions fetes.

¹⁶ La proporció d'exemples positius que estan identificats correctament pel model d'entre tots els positius reals.

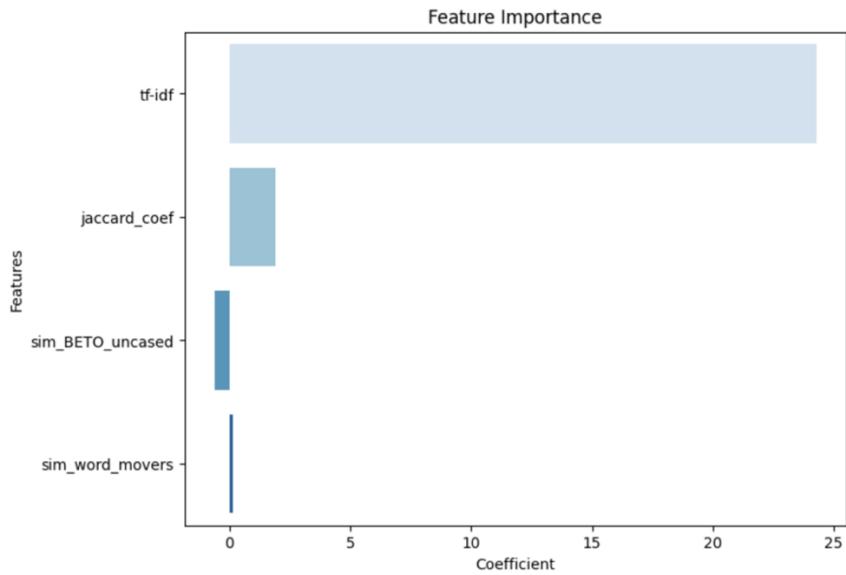


Figura 3: Importància de les característiques del classificador per parelles sobre el mateix cas

En predir sobre tots els parells d'articles, els resultats obtinguts són els que podem observar a la *Taula 13*.

ETIQUETA	Número de casos
0 - Diferent cas	1.210.997
1 - Mateix cas	48.580

Taula 13: Classificació de casos per el mateix cas

5.1.2 Agrupació per casos

Els articles d'un mateix cas s'agrupen utilitzant grafs com s'ha detallat a l'apartat 4.2.5 *Agrupació*. De tal manera que dels 14.961 articles hem aconseguit 8.198 casos distribuïts de la següent manera.

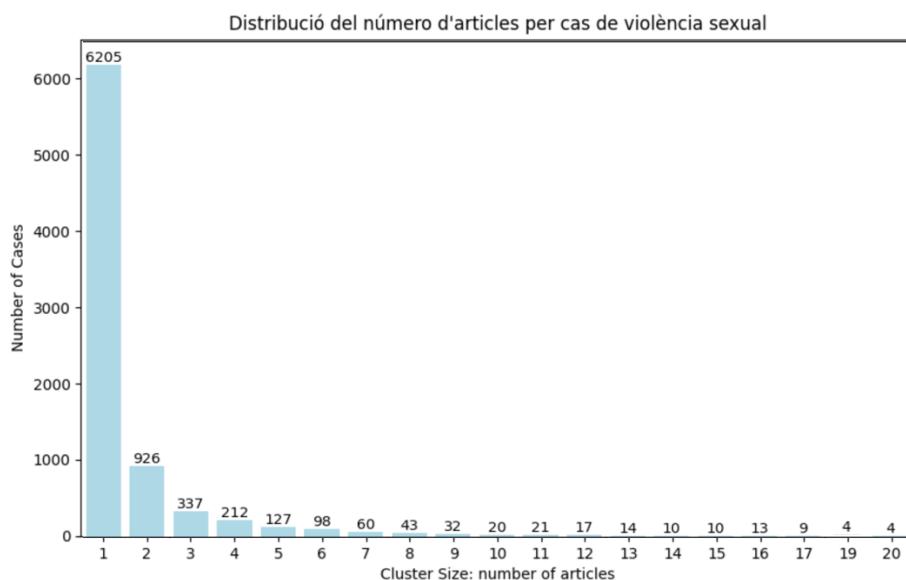


Figura 4: Distribució del número d'articles per cas de violència sexual

La Figura 4 mostra la distribució dels casos en termes del nombre d'articles que contenen. Només el 6,31% dels casos tenen una alta cobertura (mida de grup superior a 4), mentre que el 75.68% dels casos s'han explicat en un sol article.

En aquesta figura s'ha limitat els números d'articles a 20, però la nostra mostra presenta alguns *outliers*¹⁷, ja que té casos que arriben fins a 85 articles. Aquests casos tenen més soroll pel fet que utilitzem grafs connexos per classificar-los.

Per exemple, en el cas 120 que el componen 85 articles, hem comprovat que dins d'aquest clúster hi ha tres casos diferents que es produeixen exactament el mateix dia i franja horària i són molt similars.

5.2 Classificació per tipus de relació i número d'agressors

5.2.1 Relació Víctima – Aggressor

El model usat per classificar el tipus de relació entre la víctima i l'assaltant presenta la següent matriu de confusió on podem observar els TP, TN, FP i FN¹⁸ de cada etiqueta.

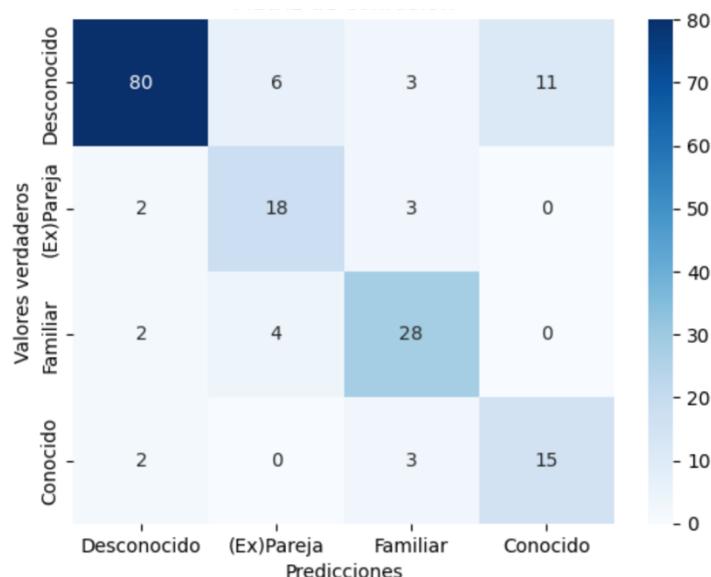


Figura 5: Matriu de confusió del model classificador víctima-agressor

A continuació, hem fet servir aquestes mètriques per obtenir la *precision*, el *recall*, *F1-score* i el *support*.

¹⁷ Un outlier és una observació anormal i extrema en una mostra estadística.

¹⁸ Sigles en anglès de: TP (veritables positius), TN (veritables negatius), FP (fals positiu), FN (fals negatiu)

Taula 14: Rendiment del model de classificació per a detectar la relació víctima-assaltant

LABELS	precision	recall	F1-score	support
Desconegut	0.93	0.80	0.86	100
(Ex)Parella	0.64	0.78	0.71	23
Familiar	0.76	0.82	0.79	34
Conegut	0.58	0.75	0.65	20

L'*F1-score* per a les categories Desconegut, Parella i Familiar és més alt que 0.7, el que denota un rendiment positiu a diferència de la classe Conegut que té molts FP que es classifiquen com a classe Desconegut, ja que el llindar entre conejut i desconegut no és trivial. La *precision* i el *recall* ens ajuden a entendre el tipus d'errors que cada característica és més probable que contingui en els resultats i a reconèixer el rang fiable de cadascun d'ells. Una baixa *precision* denota molts valors FP, mentre que un *recall* baix significa un nombre alt de FN.

En aquest cas estem considerant el model com un classificador multiclass. Ara volem analitzar el model com si aquest es tractés de diversos classificadors binaris i amb aquests, analitzar l'encert en la predicció d'esdeveniments utilitzant la mètrica AUC. Aquesta proporciona un valor numèric que representa la qualitat global del model mesurant així la probabilitat que classifiqui una mostra positiva més alta que una negativa i es defineix com l'àrea de sota la corba ROC¹⁹.

En aquests classificadors binaris, apliquem l'enfocament "un contra tots" (one-vs-all) per a problemes de classificació amb múltiples classes. Això implica que es calcula i es traca la corba ROC per a cada classe individualment, considerant aquesta classe com a positiva i totes les altres com a negatives.

Amb l'objectiu d'avaluar el model, hem generat diverses corbes ROC que es poden visualitzar amb major claredat a l'*ANNEX 2*. A més, a la *Taula 15* es mostra els resultats numèrics corresponents.

Taula 15: Valors de AUC en funció de la classe

ROC CURVE	VALOR AUC
desconegut (1) – Altres (0)	0.91
Ex/parella (1) – Altres (0)	0.93

¹⁹ La corba ROC (Receiver Operating Characteristic) és una representació gràfica del rendiment d'un model de classificació binària.

Familiar (1) – Altres (0)	0.96
Conegut (1) – Altres (0)	0.92
<i>Macro-Average de tots</i>	0.93

El classificador mostra un rendiment favorable en les classificacions individuals i en general, el que indica una capacitat efectiva per a distingir entre les diferents classes en el problema de classificació multiclasse.

Un cop el model està preparat, prediem la relació víctima-assaltant de tots els articles per tal de comparar-los amb la informació proporcionada per la *Macroencuesta 2019* (Ministerio de Igualdad de España, 2020).

Taula 16: Nombre d'articles corresponents per tipus de vincle entre la víctima i l'agressor

Relació Víctima Agressor	Macroencuesta 2019	Nombre d'articles	Percentatge d'articles
Desconeget	10.86%	9183	61,38%
Conegut	14.03%	3266	21,83%
Familiar	6.00%	1477	9,87%
(Ex)Parella	72.20%	1035	6,92%

Es pot observar que la majoria dels delictes de violència sexual ocorren quan hi ha o hi havia una relació entre la víctima i l'agressor. No obstant això, els mitjans de comunicació mostren una realitat diferent.

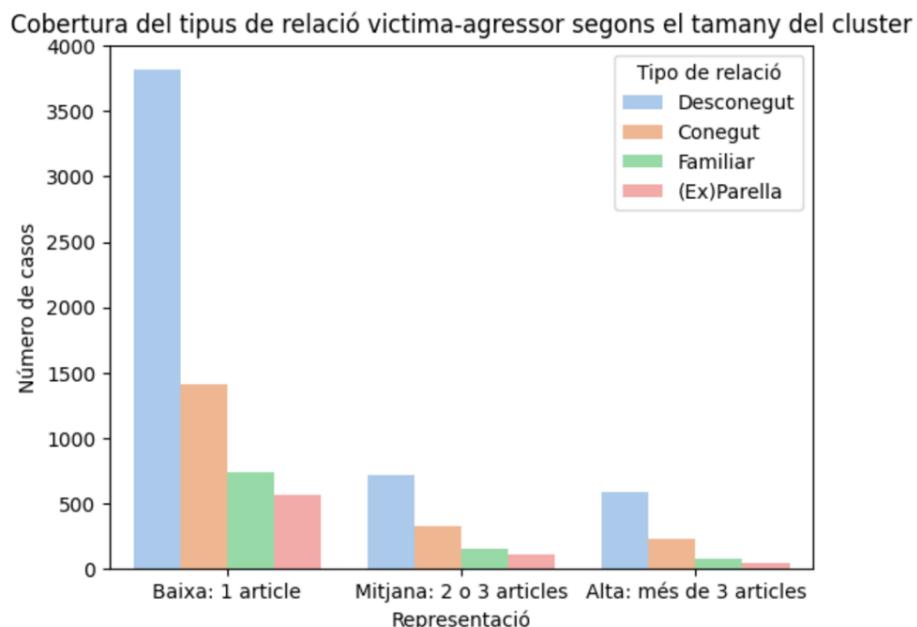


Figura 6: Representació del tipus de relació entre la víctima i l'agressor en funció de la mida del clúster

D'altra banda, podem veure que en escassos casos, quan el vincle és "familiar" o de "(ex)parella", el cas no està molt cobert pels mitjans de comunicació.

5.2.2 Número d'agressors

El model utilitzat per classificar el número d'autors involucrats en l'agressió presenta la següent matriu de confusió on podem observar els TP, TN, FP i FN de cada etiqueta.

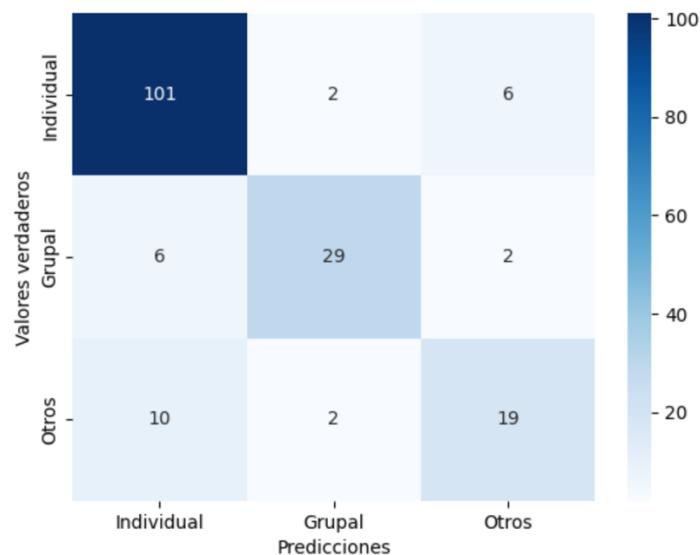


Figura 7: Matriu de confusió del model classificador del número d'agressors

Taula 17: Rendiment del model de classificació per a detectar el número d'agressors

LABELS	precision	recall	F1-score	support
Individual	0.86	0.93	0.89	109
Grupal	0.88	0.78	0.83	37
Altres	0.70	0.61	0.66	31

L'*F1-score* per a les categories Individual i Grupal és més alt que 0.8, el que denota un rendiment molt positiu a diferència de la classe Altres que té molts FP. Aquesta *F1-score* baixa és lògica atès que la classe Altres engloba tant casos que no s'especifica el nombre d'agressors com articles informatius sobre noves lleis, o canvis d'aquestes, etc.

Com hem explicat amb l'anterior classificador, hem aplicat el mateix procediment de cara a avaluar el model, generant diverses corbes ROC que es poden visualitzar amb major claredat a l'*ANNEX 2*. A més, la *Taula 18* ens mostra els resultats numèrics corresponents.

Taula 18: Valors de AUC en funció de la classe

ROC CURVE	VALOR AUC
Individual (1) – Altres (0)	0.91
Grupal (1) – Altres (0)	0.91
Altres (1) – Altres (0)	0.85
<i>Macro-Average de tots</i>	0.89

El classificador mostra un rendiment satisfactori en les classificacions individuals, lleugerament inferior per a la classe "Altres", el que indica en general una capacitat efectiva per a distingir entre les diferents classes en el problema de classificació.

Un cop el model està preparat, prediem el número d'agressors de tots els articles per tal de comparar-los amb la informació proporcionada per la *Macroencuesta 2019* (Ministerio de Igualdad de España, 2020).

Taula 19: Nombre d'articles corresponents per tipus d'agressió segons el nombre d'agressors

Número d'autors involucrat	Macroencuesta 2019	Nombre d'articles	Percentatge d'articles
Individual	87.96%	9202	61,50%
Grupal	12.04%	1657	11,07%
Altres	-	4102	27,41%

Es pot observar que la majoria dels delictes de violència sexual són casos individuals.

Cobertura de casos d'agresions grupals i individuals segons el tamany del cluster

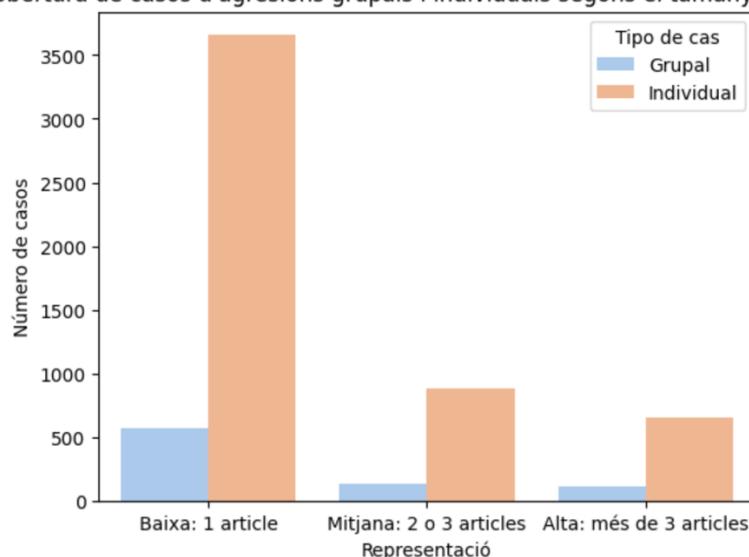


Figura 8: Representació del tipus de relació entre la víctima i l'agressor en funció de la mida del clúster

D'altra banda, a la *Figura 8*, podem observar que sempre hi ha una tendència cap a una baixa cobertura de casos pels mitjans de comunicació.

5.3 Anàlisis del vocabulari

Com hem explicat a l'apartat 4.4 *Anàlisis del vocabulari*, distingim 4 categories gramaticals i per cadascuna d'elles examinem el vocabulari emprat en els casos en funció del número d'agressors i del tipus de relació entre la víctima i l'agressor.

El principal a entendre primer de tot és la llegenda presentada a dalt de tots els gràfics on es poden veure representat les següents combinacions de símbols per a saber la contribució de cada paraula en el text:

- $+ \uparrow$: Paraula relativament positiva que s'utilitza més sovint en el text
- $+ \downarrow$: Paraula relativament positiva que s'utilitza menys sovint en el text
- $- \uparrow$: Paraula relativament negativa que s'utilitza més sovint en el text
- $- \downarrow$: Paraula relativament negativa que s'utilitza menys sovint en el text
- Δ : La puntuació de la paraula en el text és alta
- ∇ : La puntuació de la paraula en el text és baixa

Podem veure els resultats de tots els gràfics generats a l'*ANNEX 3*.

Per tal d'entendre millor els resultats obtinguts, a la següent *Figura 9*, es presenta una comparació del vocabulari utilitzat en textos d'agressions grupals (part esquerra del gràfic) o individuals (part dreta del gràfic) en funció dels verbs usats. Es pot observar que els verbs fets servir per descriure els casos de violència grupal (*agresion, violar, agredir, violaron, agredieron*) són més agressius en comparació amb els usats en un cas d'un sol agressor, on el verb més freqüent és *abusar*. A partir d'aquests gràfics, podem extreure diverses conclusions.

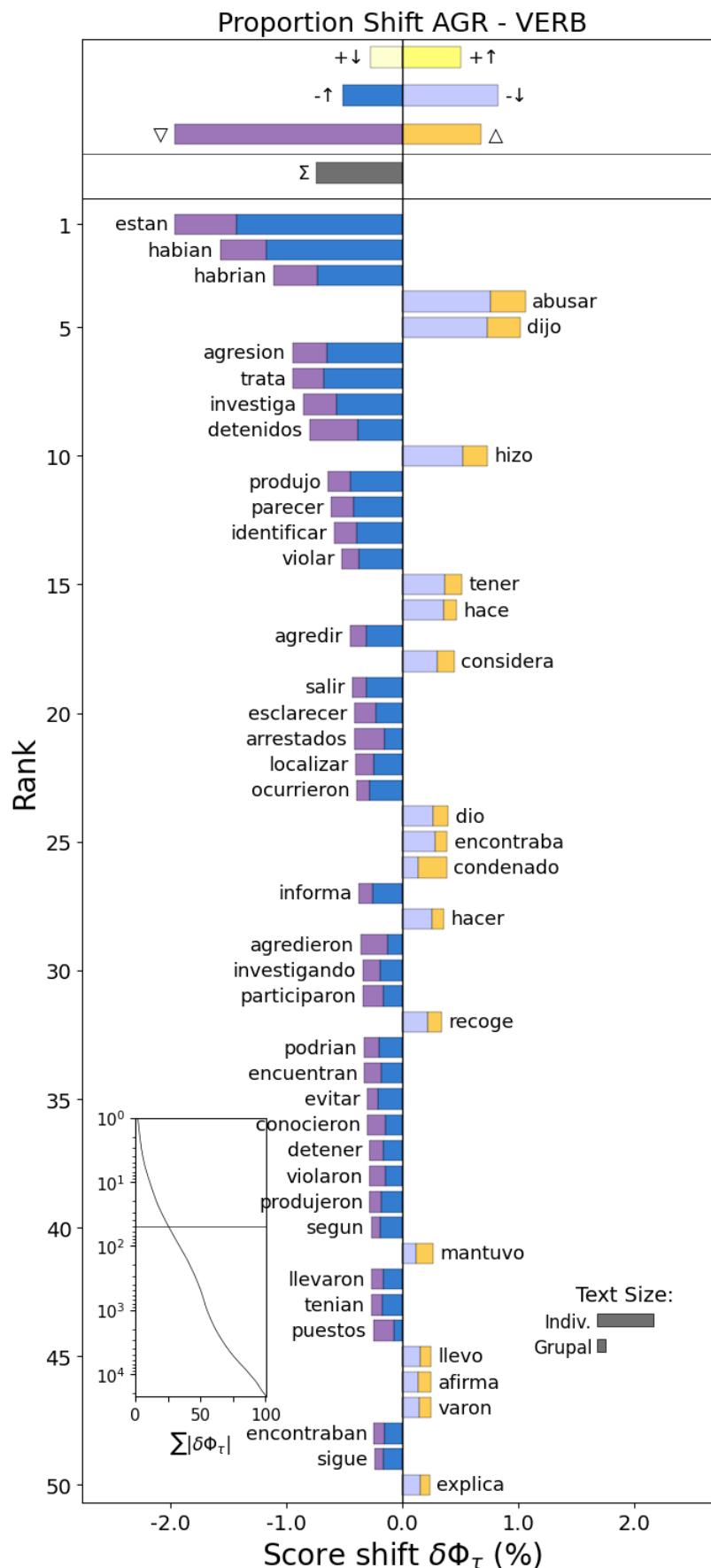


Figura 9: Comparació entre el vocabulari emprat en textos d'agressions Grupals o Individuals en funció dels verbs utilitzats

6. DISCUSSIÓ I CONCLUSIONS

En aquest apartat s'exposen les conclusions del projecte, basades en els objectius plantejats anteriorment a l'apartat 1.2 i en els resultats descrits a l'apartat anterior.

6.1 Discussió

6.1.1 Classificació per casos

Tal com hem explicat anteriorment, l'objectiu d'aquest apartat era agrupar tots els articles que parlessin sobre el mateix cas. En el següent exemple, veurem la comparació de resultats sobre dos articles classificats com a el mateix cas, i dos articles classificats com diferents casos.

Exemple 1: Característiques per parells d'articles en els dos escenaris possibles

Característiques d'un parell d'articles sobre el mateix cas

Títol i titular de l'article A²⁰: *detiene 3 jovenes acusados violar 3 hermanas estadounidenses,fuentes arrestados estudian ciudad formacion profesional 20 25 anos*

Títol i titular de l'article B²¹: *3 jovenes acusados violar 3 hermanas estadounidenses,detenido jueves tres jovenes acusados violaciones denunciadas tres hermanas estadounidenses domicil*

WMD	Jaccard	TF-IDF	BETO
0.0920	0.0687	0.8313	0.9856

Característiques d'un parell d'articles sobre diferent cas

Títol i titular de l'article A²²: *caso abuso sexual menores principe tras acuerdo privado,orden dado despues monarca realizado importante donacion organizacion benefica acuso haber abusado tenia 17 anos trama trafico sexual epstein*

Títol i titular de l'article B²³: *espanola recibido 506 denuncias abusos sexuales menores dos anos, formara parte comision investigar abusos*

WMD	Jaccard	TF-IDF	BETO
0.2212	0.0709	0.0706	0.8068

Segons la importància de les característiques presentada a la *Figura 6*, la característica més rellevant en la classificació és la *cosine similarity* del TF-IDF. En l'*Exemple 1*, podem observar que és la mètrica que més varia entre tots dos casos. Encara que aquesta

²⁰ <https://www.20minutos.es/noticia/4104199/0/detienen-a-3-jovenes-acusados-de-violar-a-3-hermanas-estadounidenses-en-murcia/>

²¹ https://www.cope.es/emisoras/region-de-murcia/murcia-provincia/murcia---san-javier/informativos-en-murcia/noticias/detenidos-jovenes-por-violacion-hermanas-estadounidenses-20200103_585455

²² <https://www.rtve.es/noticias/20220309/desestimado-abuso-sexual-menores-principe-andres-acuerdo-privado/2305960.shtml>

²³ <https://www.larazon.es/sociedad/20220311/wny5ylhckjeh7gvxttucurdxii.html>

característica té una gran influència en la predicció, la combinació de característiques ajuden al rendiment de l'algorisme per a l'avaluació de la similitud de tots dos articles. Seguidament, al considerar tots els articles que pertanyen al mateix cas com les components connexes d'aquest, amb un llindar de 0.85, la precisió a l'agrupar és prou alta per agrupar bé (*Exemple 2*) però, ens trobem amb alguns pocs casos on la mida del clúster és molt gran i aquest inclou més d'un cas, com en el següent *Exemple 3*.

Exemple 2: Agrupació de clústers correcte

Clúster id: 17 – Tamany del clúster: 2

Títol de l'article A²⁴: *difundir videos sexuales amiga consentimiento*

Títol de l'article B²⁵: *dos varones acusados difundir videos sexuales amiga consentimiento*

Exemple 3: Agrupació de clústers errònia

Clúster id: 120 – Tamany del clúster: 85

Titular de l'article A²⁶: *madre detenido violacion echo abusar hermana 7 anos*

Titular de l'article B²⁷: *detenido brutal violacion acusado agredir sexualmente hermana siete anos*

Titular de l'article C²⁸: *arresto producido tras cinco meses investigaciones*

Titular de l'article D²⁹: *desquada detenido madrugada joven 20 anos supuesto autor agresion sexual menor tras compleja investigacion cinco meses*

En els casos en què la grandària del clúster és extremadament gran, com en el cas de l'*Exemple 3*, en revisar els casos que en formen part, aquest presenta subclústers (podem observar com els articles A i B pertanyen al mateix cas, així com ho fan el C i D). Per abordar això, podríem augmentar el llindar, però això generaria molts clústers de grandària 1 quan en realitat haurien de pertànyer a un de grandària 2, per exemple.

És evident que la cobertura mitjana que tenen els casos sobre violència sexual és molt baixa. Com podem veure a la *Figura 7*, més del 90% dels casos no s'expliquen en més de tres articles entre tots els mitjans de comunicació.

²⁴ <https://www.elperiodico.com/es/sociedad/20210903/detenidos-granada-difundir-videos-sexuales-12041354>

²⁵ <https://www.20minutos.es/noticia/4809479/0/detenidos-por-difundir-videos-sexuales-de-una-amiga-sin-su-consentimiento/>

²⁶ https://www.elconfidencial.com/espana/cataluna/2022-04-22/violacion-igualada-detenido-denunciado-por-madre-por-agredir-sexualmente-hermana-7-anos_3412690/

²⁷ https://www.lasexta.com/noticias/sociedad/detenido-brutal-violacion-igualada-fue-acusado-agredir-sexualmente-hermana-pequena_2022042262627f8240d8080001a42f11.html

²⁸ <https://cadenaser.com/2022/04/21/detenido-un-hombre-por-la-brutal-agresion-sexual-a-una-menor-en-igualada-en-2021/>

²⁹ https://www.elconfidencial.com/espana/cataluna/2022-04-21/violacion-igualada-menor-pistas-huellas-adn-victima-ropa-detenido_3412002/

6.1.2 Anàlisi del vocabulari

Són clars els mites, estigmes i estereotips presents en la societat sobre els casos de violència sexual. Entre aquests, podríem dir que la gran majoria d'agressors són desconeguts, que les agressions en grup són relativament freqüents i extremadament violentes, és més probable que les agressions tinguin lloc durant la nit i sota els efectes de l'alcohol, etc. Mitjançant l'anàlisi efectuada a l'apartat 4.4 i els resultats obtinguts a l'apartat 5.3, podem veure reflectits de manera conjunta aquests estereotips i mites. Per una millor comprensió de les conclusions que s'exposen a continuació, recomano consultar l'*ANNEX 3*.

En la comparació del vocabulari entre un cas d'agressió grupal amb un d'individual, veiem certes tendències:

- En ambdós casos s'utilitzen paraules com “*presunto/a*”, “*supuestamente*”. Posant en dubte el cas, donant a entendre que no està demostrat.
- Els casos d'agressions grupals, presenten paraules amb molt més detall.
- Els verbs emprats per a les agressions grupals, són més agressius (*agredir/agresion/agredieron, violar/violaron*) en canvi, en el cas d'un sol agressor el verb més usat és “*abusar*”.
- Normalment, es fan servir paraules com: *brutal, grupal, múltiple...*

Quan l'agressor és o ha sigut la parella *vs* un desconegut observem que:

- Se segueixen presentant les paraules “*presuntos/as*” en ambdós casos.
- Com a paraules més comunes quan es descriu el cas de parella, són “*domicilio*”, “*vivienda*”, “*casa*”. Això ens fa veure que les agressions sempre es duen a terme dins del domicili.
- Els verbs usats per a les agressions de parella són més agressius (*violar, matar, cadàver, agredir*) i deixen entreveure com si fos un cas excepcional portat a l'extrem. Quan hem pogut corroborar que no és així, ja que, a la *Macroencuesta 2019* es veu com més del 70% de les agressions són dutes a terme per la (ex)parella.

I finalment, quan l'agressor ha sigut un familiar de la víctima *vs* un desconegut:

- En el cas d'un familiar, la víctima sempre se la presenta com a “*menor*” o “*nina*”.
- S'utilitzen les paraules “*continuado*” i “*ocasiones*”, mostrant així que és habitual que una agressió per part d'un familiar no sigui una agressió esporàdica.

- I en coincidència amb el cas de parella, per familiar també s'usen les paraules “domicilio”, “vivienda” i “casa”.
- En els casos on l'agressor és un desconegut, s'utilitzen molt les paraules “sociales”, “local”, “publico” i a més a més se suma la paraula “atenuante”.

6.2 Conclusions i futurs estudis

Aquest estudi s'ha fet amb els objectius de desenvolupar un model automatitzat capaç d'identificar i agrupar els articles que tracten sobre el mateix cas per veure en quina proporció s'aborden els casos als mitjans, crear un classificador multiclasse basat en models de processament del llenguatge natural per determinar la relació existent entre l'assaltant i la víctima, crear un classificador que permeti detectar i determinar si hi ha un o més agressors implicats en un cas específic, analitzar i comparar el llenguatge emprat en cadascun d'aquests dos casos i comparar amb les estadístiques nacionals oficials si els mitjans representen amb la mateixa mesura tots els tipus de cas i de quina manera ho fan.

La primera conclusió que podem extreure és que els mitjans de comunicació presenten els casos de violència sexual com un esdeveniment allunyat de la realitat i en situacions molt concretes. Aquesta conclusió es pot confirmar a través de la sobre representació dels delictes causats per desconeguts a diferència de la sota representació dels articles comesos per parelles o exparelles de la víctima. Amb la poca representació de la realitat, el que aconsegueixen els mitjans de comunicació és desviar l'atenció cap a les situacions estereotipades i reforçar encara més els mites existents.

Per altra banda, és una evidència que el vocabulari utilitzat per a descriure diferents casos varia en funció del tipus de cas que s'estigui descrivint. L'únic factor que tenen en comú tots els casos és les paraules de dubte com “presumptament” i “suposadament” oferint-li així al lector el dret a dubte sobre l'agressor. A més a més, en els casos que l'autor de l'agressió és un familiar o la (ex) parella de la víctima, el llenguatge usat és més agressiu i situa als casos en fets extremos, com si aquests autors no fossin persones normals que ens envolten a la vida quotidiana, sinó casos aïllats de la normalitat. Això es pot demostrar més amb el fet que la cobertura d'aquest tipus de casos és molt escassa.

Totes aquestes conclusions recolzen la idea que la cobertura de la violència sexual en els mitjans de comunicació espanyols no reflecteix la realitat d'aquest tema tan delicat, la qual cosa té conseqüències en la forma en què la població percep la violència sexual.

Amb la finalitat de combatre el problema de la violència sexual, els mitjans de comunicació han de proporcionar exemples precisos de violació que no s'ajustin a les nocions preconcebudes ni s'ajustin als mites. Només a través d'això, els mitjans de comunicació poden començar a abordar els problemes socials amb la violència sexual.

De cara a possibles futurs estudis, podria ser interessant:

- Analitzar més en profunditat el vocabulari emprat en cada cas d'agressió sexual tenint en compte més factors; com per exemple, el lloc de l'agressió, si l'article ha estat escrit per un home o per una dona, etc.
- Analitzar més en detall els casos més mediàtics, ja sigui per número de mitjans que parlen sobre el cas o per la durada d'aquest en els mitjans, i veure si canvia la manera d'expressar-se envers els casos menys rellevants o si durant el procés del cas d'aquest mateix, el vocabulari fet servir varia.

7. BIBLIOGRAFIA

A. G. D'Sa, I. Illina and D. Fohr, "BERT and fastText Embeddings for Automatic Detection of Toxic Speech," *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*, Tunis, Tunisia, 2020, pp. 1-5, doi: 10.1109/OCTA49274.2020.9151853

A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, (2016). "FastText.zip: Compressing Text Classification Models", ArXiv Prepr. ArXiv161203651

Almeida, F., Geraldo, X. (2023). Word Embeddings: A Survey
<https://arxiv.org/abs/1901.09069v2>

Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, Edward A. Fox 2021 Natural Language Processing Advancements By Deep Learning: A Survey arXiv:2003.01200

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. ACM press, Vol. 463.

Bahrdwaj, Saksham; Saxena, N. (2019). Word Mover's Distance for Text Similarity. Towards Data Science.

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>

Cavnar, W. B., Trenkle, J. M., et al. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Vol. 161175. Citeseer.

Devlin et al., NAACL (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding <https://aclanthology.org/N19-1423>

ESTRICH, Susan (1987). Real Rape. How the Legal System Victimizes Women Who Say No. Cambridge: Harvard University Press.

Fitzpatrick, N. (2018). Media Manipulation 2.0: The Impact of Social Media on News, Competition, and Accuracy. *Athens Journal of Mass Media and Communications*, 4(1), 45–62. <https://doi.org/10.30958/ajmmc.4.1.3>

Gallagher, R. J. (2020). SHIFTERATOR. <https://shifterator.readthedocs.io>

Gallagher, R. J., Morgan R. Frank, Lewis Mitchell, Aaron J. Schwartz, Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds. “Generalized Word Shift Graphs: A Method for Visualizing and Explaining Pairwise Comparisons Between Texts.”” *EPJ Data Science* 10, no. 4 (2021).

Instituto Nacional Estadístico (2023) https://www.ine.es/prensa/evdvg_2022.pdf

Kim, SW., Gil, JM (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Hum. Cent. Comput. Inf. Sci.* 9, 30 <https://doi.org/10.1186/s13673-019-0192-7>

Kleinbaum, D.G., Klein, M. (2010) <https://doi.org/10.1007/978-1-4419-1742-3>

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2005). Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.*, 30, 25–36.

Lei Xu et al., (2019). “Modeling tabular data using conditional GAN”. In: *Advances in Neural Information Processing Systems*. <https://arxiv.org/pdf/1907.00503.pdf>

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Trans. Intell. Syst. Technol.* 13(2), Article 31, 41 pages. <https://doi.org/10.1145/3495162>

Mikolov, T., et al. (2013). Efficient estimation of word representations in vector space. In Proc. ICLR, 2013. <http://arxiv.org/abs/1301.3781>

Ministerio de Igualdad, (2018). PERCEPCIÓN SOCIAL DE LA VIOLENCIA SEXUAL https://violenciagenero.igualdad.gob.es/violenciaEnCifras/estudios/colecciones/pdf/Libr_o_25_Violencia_Sexual.pdf

Ministerio de Igualdad de España (2020). Resumen ejecutivo de la Macroencuesta de Violencia contra la Mujer 2019.

<https://violenciagenero.igualdad.gob.es/violenciaEnCifras/macroencuesta2015/Macroencuesta2019/home.htm>

Ministerio de Interior de España (2019). Informe sobre delitos contra la libertad e indemnidad sexual en España.

<https://estadisticasdecriminalidad.ses.mir.es/publico/portalestadistico/dam/jcr:4d37cea3-eafd-49e0-9ae8-82aa73b52e2a/INFORME%20DELITOS>

Morgan, S. (2018). Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, 3(1), 39–43.
<https://doi.org/10.1080/23738871.2018.1462395>

Newman, N. et al. (2020). Reuters Institute Digital News Report 2020
https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard coefficient for keywords similarity. In Proceedings of the International Multiconference of Engineers and Computer Scientists, Vol. 1, 380–384.

Partalidou, E., et al. (2019). DESIGN AND IMPLEMENTATION OF AN OPEN SOURCE GREEK POS TAGGER AND ENTITY RECOGNIZER USING SPACY.
<https://arxiv.org/pdf/1912.10162.pdf>

Partalidou, E., et al. (2019). DESIGN AND IMPLEMENTATION OF AN OPEN SOURCE GREEK POS TAGGER AND ENTITY RECOGNIZER USING SPACY.
<https://arxiv.org/pdf/1912.10162.pdf>

Ryoma Sato, Yamada, M., & Kashima, H. (2022). Re-evaluating Word Mover's Distance v3. arXiv:2105.14403

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 27-64.
<https://doi.org/10.1016/j.cosrev.2007.05.001>

Sreedharan, C., Thorsen, E., & UNESCO Office in New Delhi. (2021). Sexual violence and the news media: issues, challenges and guidelines for journalists in India.

Sriram, S. (2020). An Evaluation of Text Representation Techniques for Fake News Detection 85 Using : TF-IDF, Word Embeddings, Sentence Embeddings with Linear Support Vector Machine. *Arrow@TU Dublin*. <https://doi.org/10.21427/5519-h979>

Walton, N. A. (2020). Myths, messages, and the media: The media's role in perpetuating sexual harrassment stereotypes.

World Health Organization (2013). *Global and Regional Estimates of Violence against Women: Prevalence and health effects of intimate partner violence and non-partner sexual violence*. Geneva: World Health Organization.

Yin Zhang, Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics, 1(1–4), 43–52.

ANNEX 1. Dades i codi

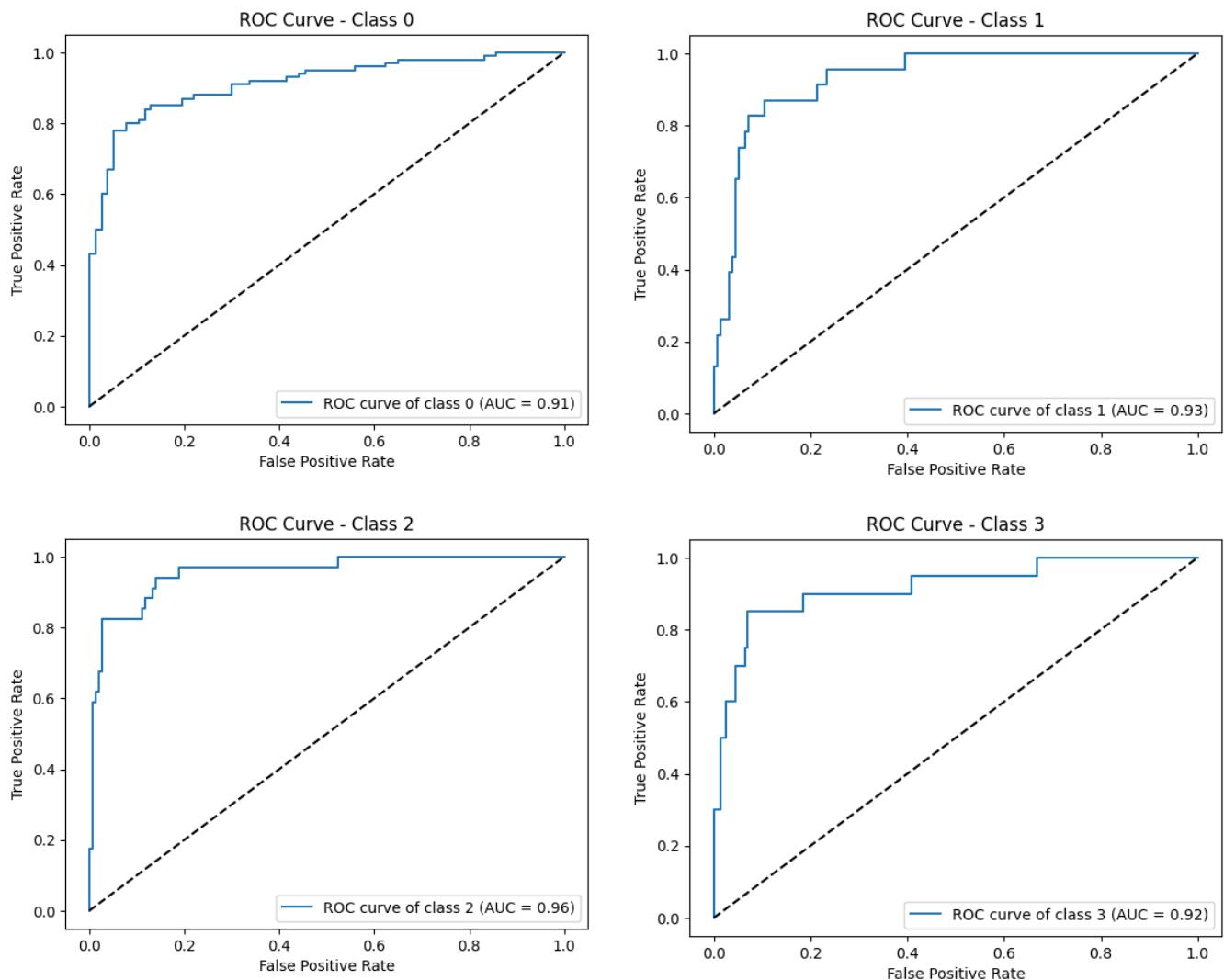
Les dades i el codi utilitzat per la realització d'aquest treball es poden trobar a:

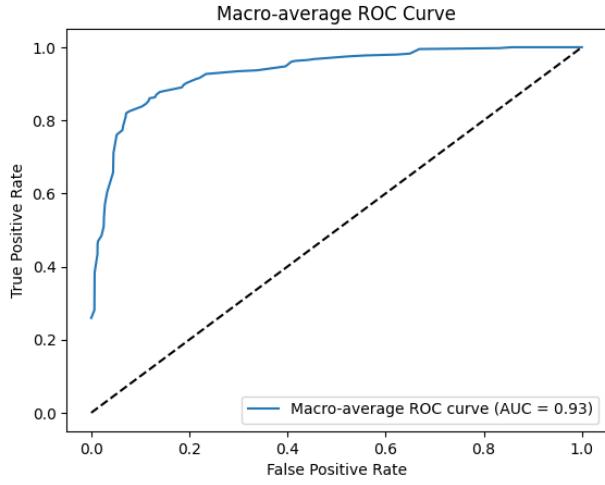
<https://github.com/martagraciavalles/Aprendentatge-autom-tic-i-la-viol-ncia-sexual-a-espanya>

ANNEX 2. Resultats classificadors de tipus de relació aggressor-víctima i número d'agressors involucrats en el cas

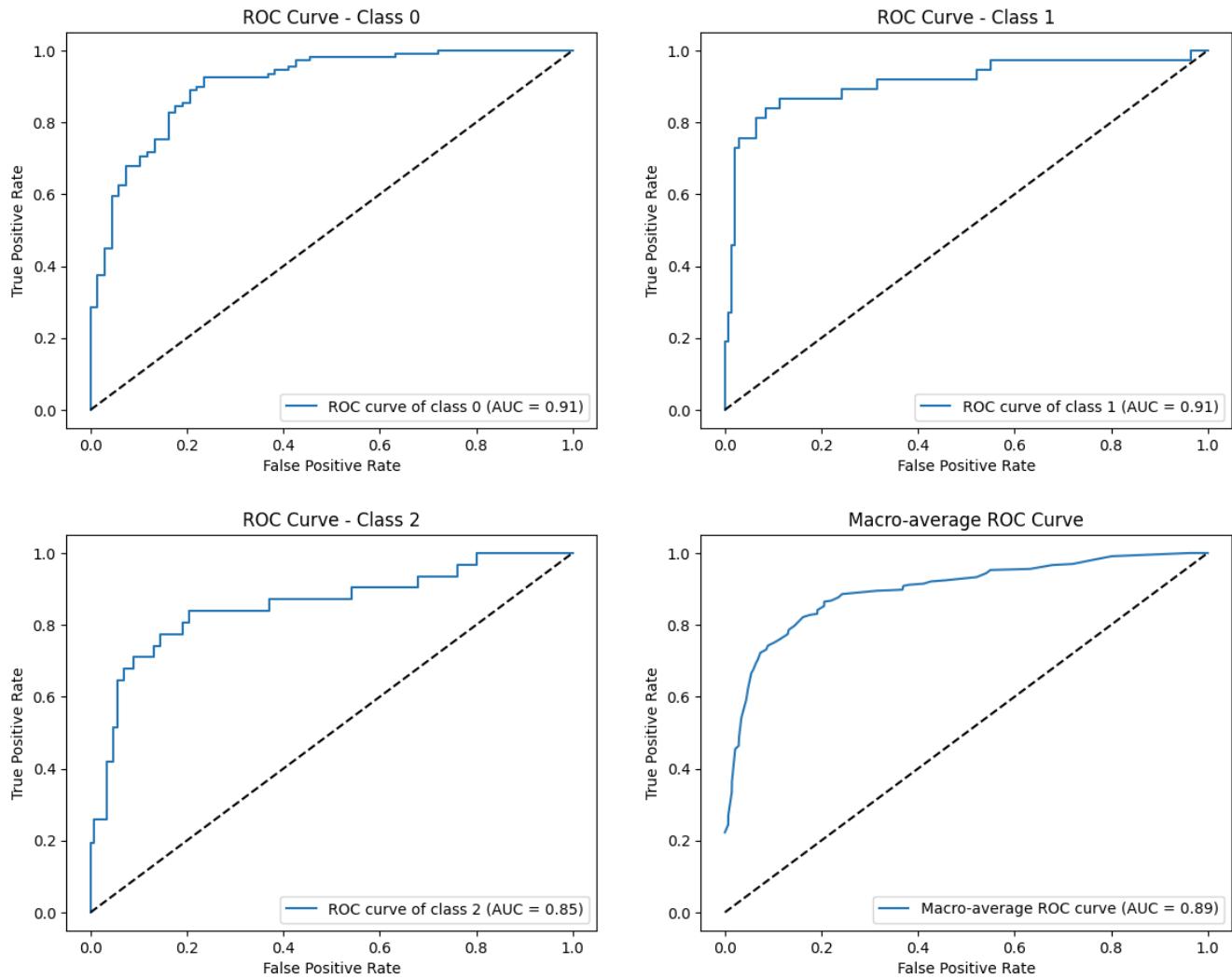
A continuació es presenten les figures que mostren els resultats de les corbes ROC per a cada classificador binari descrit a l'apartat 5.2 *Classificació per tipus de relació i número d'agressors*.

En primer lloc, es presenten les gràfiques representatives dels classificadors binaris en funció de la relació entre l'agressor i la víctima. En cadascuna d'elles, considerem que la classe que avaluem és la positiva, mentre que totes les altres es consideren negatives. També es mostra la gràfica de la mitjana de totes les classes (*Macro-average ROC curve*). Aclarir que la classe 0 es tracta de la relació ‘Desconeugut’ entre la víctima i l’agressor, la 1 és ‘(Ex)Parella’, la 2 és ‘Familiar’ i la 3 és ‘Conegut’.





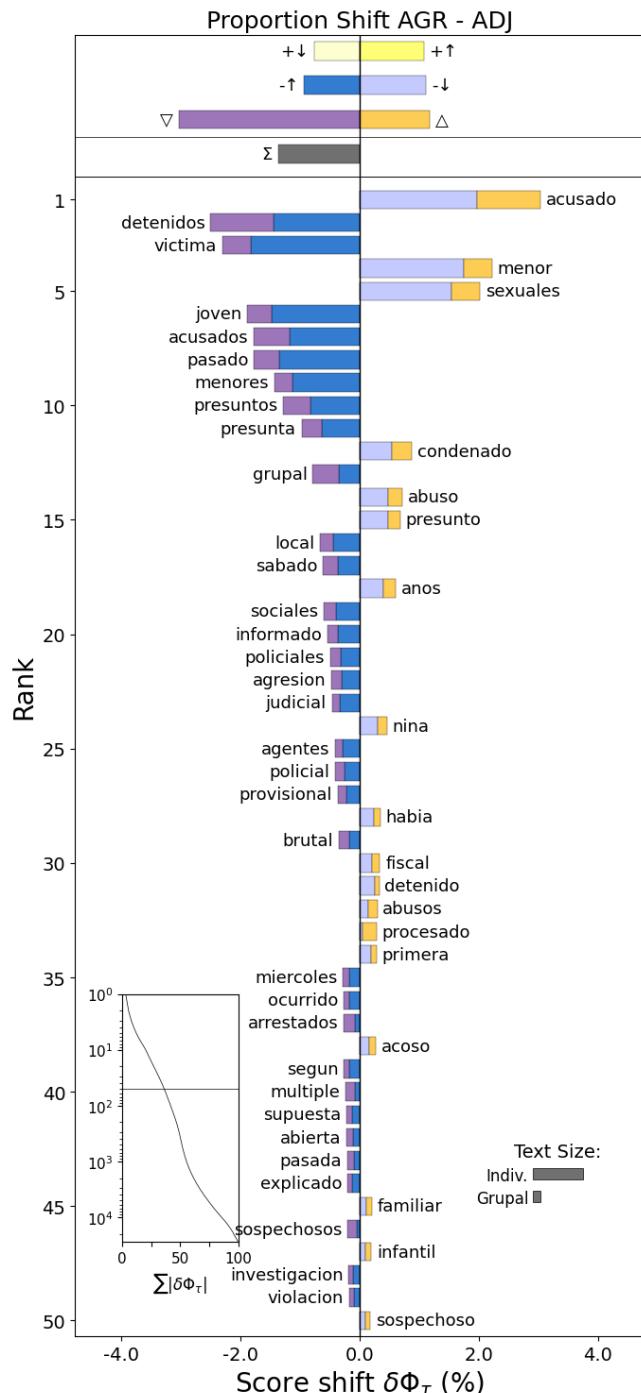
Finalment, es mostren les gràfiques representatives dels classificadors binaris en funció del número d'agressors involucrats ens el cas. En cadascuna d'elles, considerem que la classe que avaluem és la positiva, mentre que totes les altres es consideren negatives. També es mostra la gràfica de la mitjana de totes les classes (*Macro-average ROC curve*).



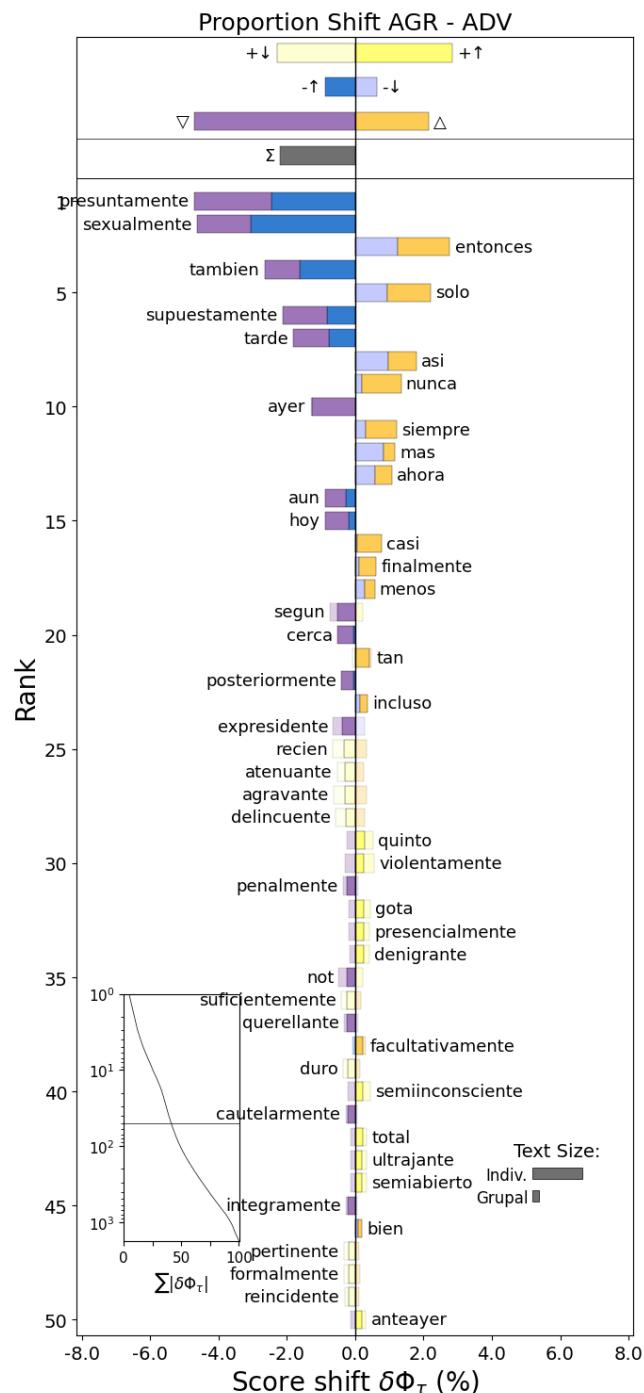
ANNEX 3. Resultats anàlisi del vocabulari

A continuació es presenten les figures que mostren els resultats visuals per a cada cas i categoria determinada a l'apartat 5.3 *Anàlisi del vocabulari* descrit anteriorment.

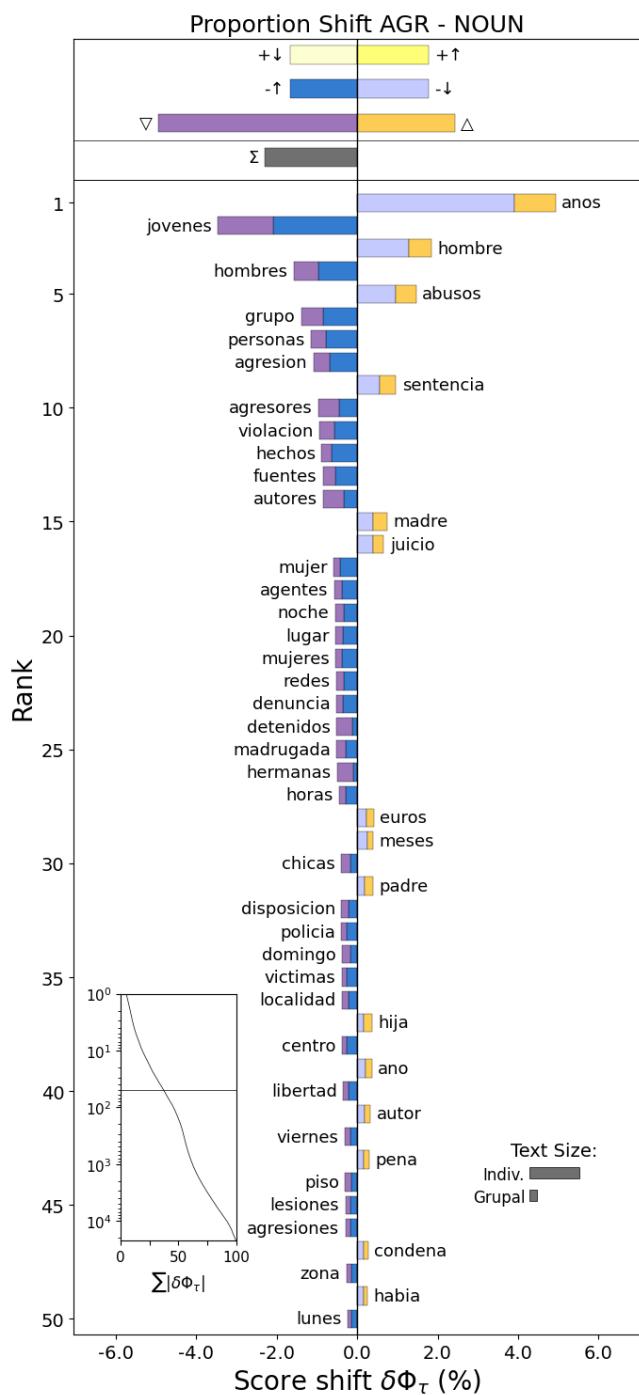
En primer lloc, es mostra la comparació visual del vocabulari emprat en els textos amb un únic agressor (part dreta de tots dos gràfics) en contrast amb una agressió grupal (part esquerra dels gràfics), segons la categoria gramatical.



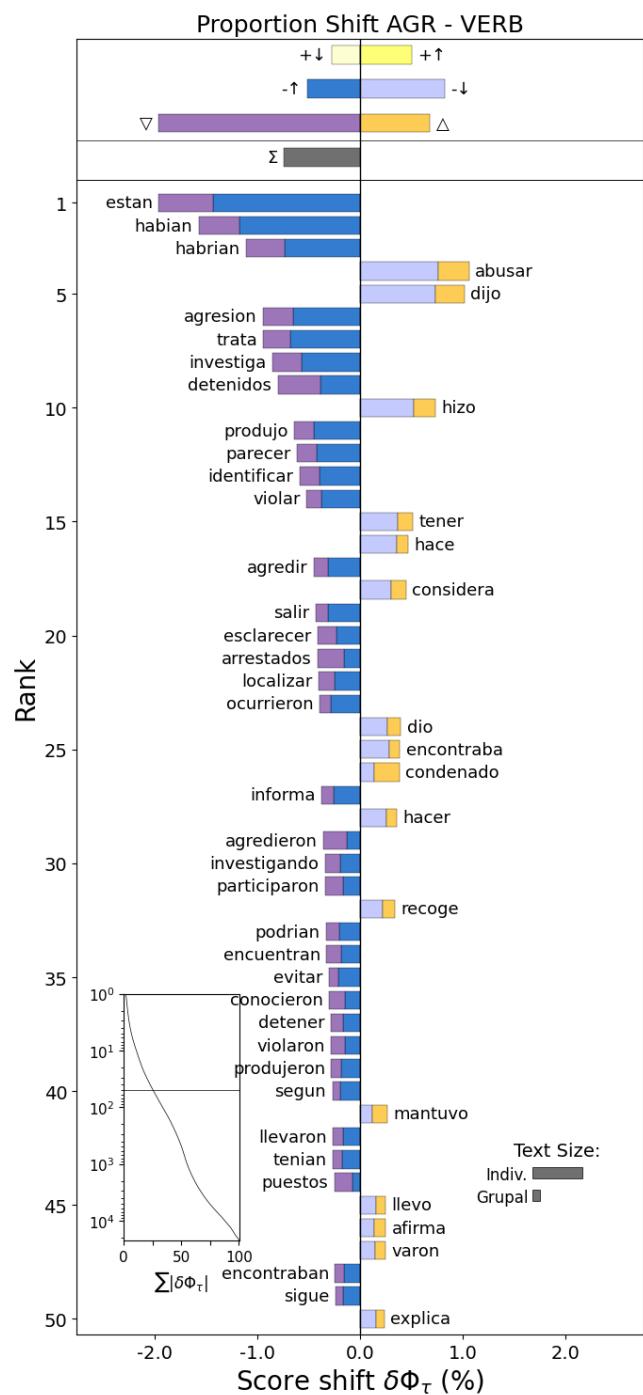
Comparació entre el vocabulari emprat en textos d'agressions Grupals i Individuals en funció dels *adjectius* utilitzats



Comparació entre el vocabulari emprat en textos d'agressions Grupals i Individuals en funció dels *adverbis* utilitzats

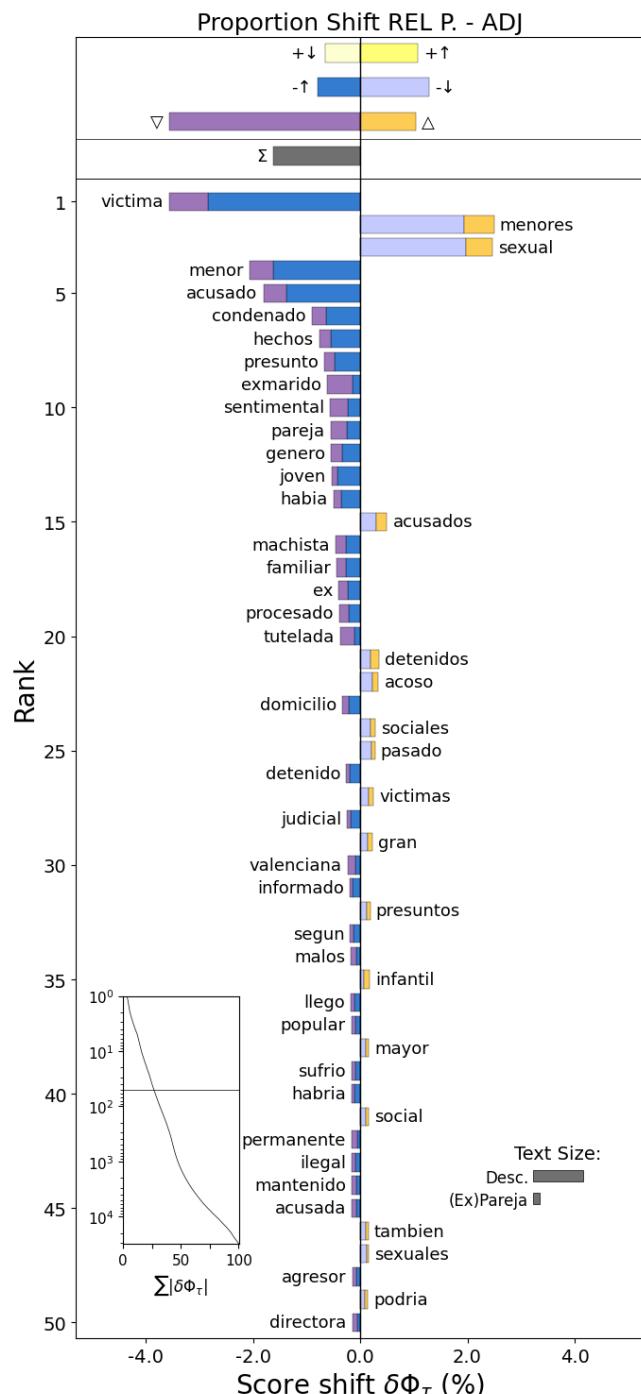


Comparació entre el vocabulari emprat en textos d'agressions Grupals o Individuals en funció dels noms utilitzats

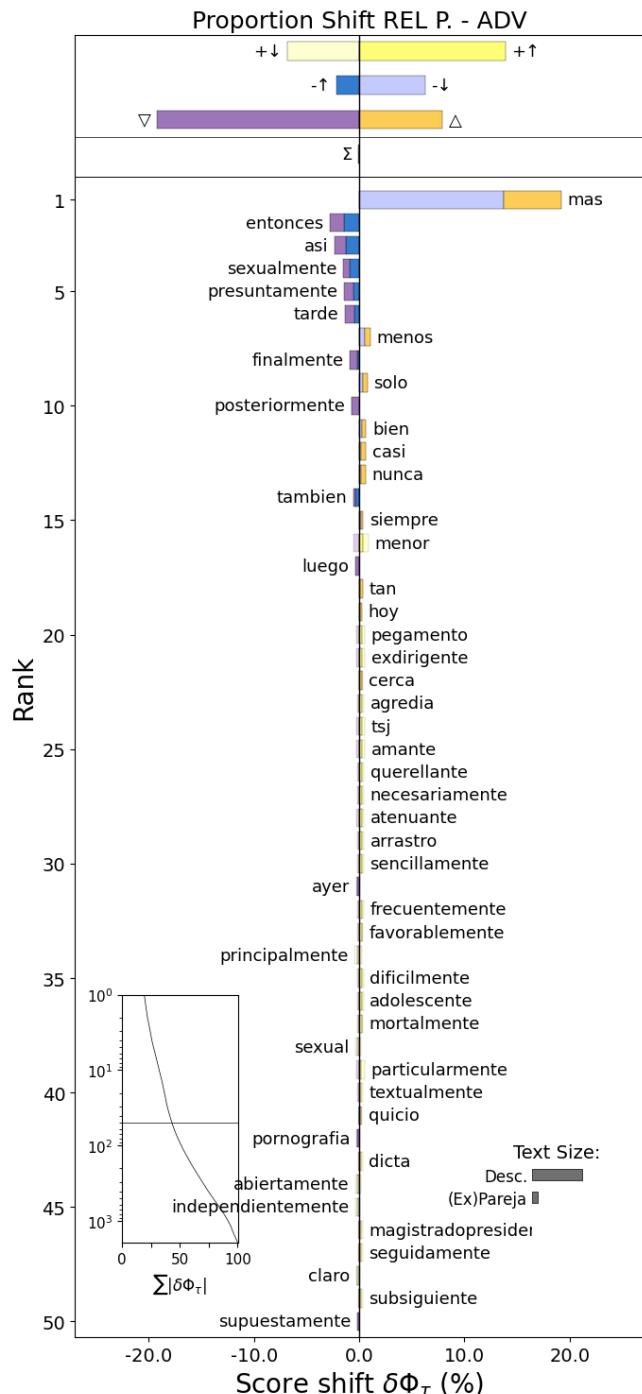


Comparació entre el vocabulari emprat en textos d'agressions Grupals o Individuals en funció dels verbs utilitzats

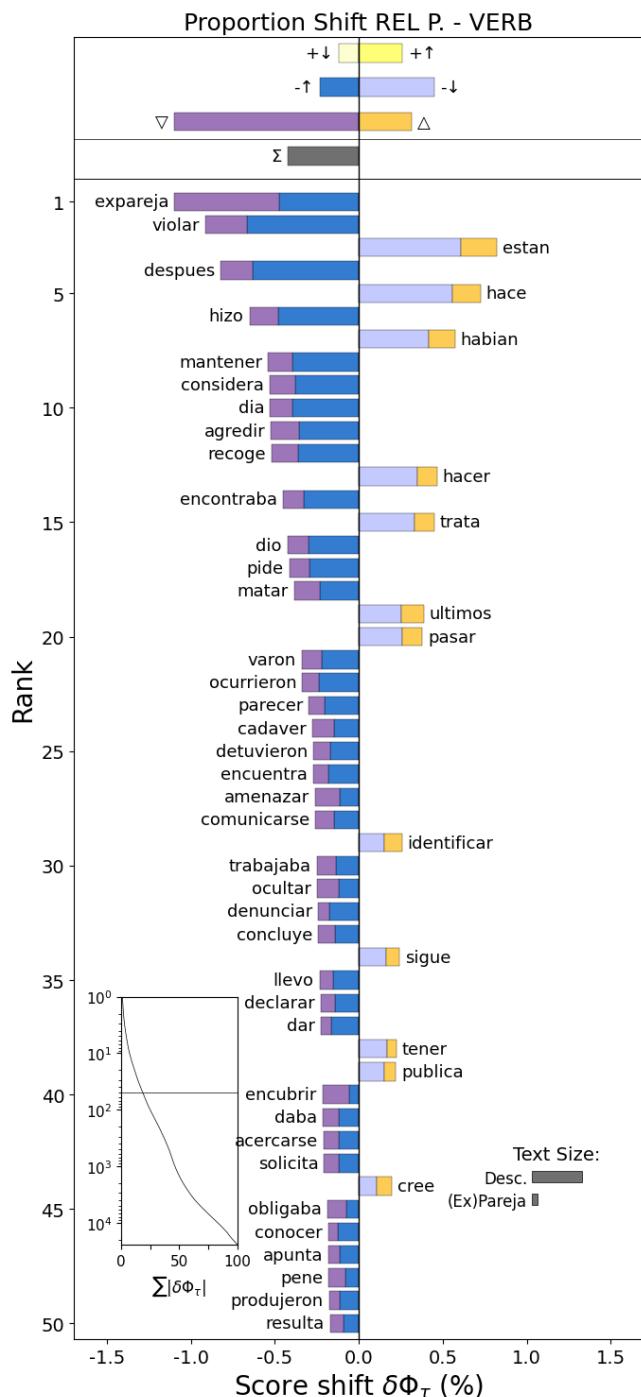
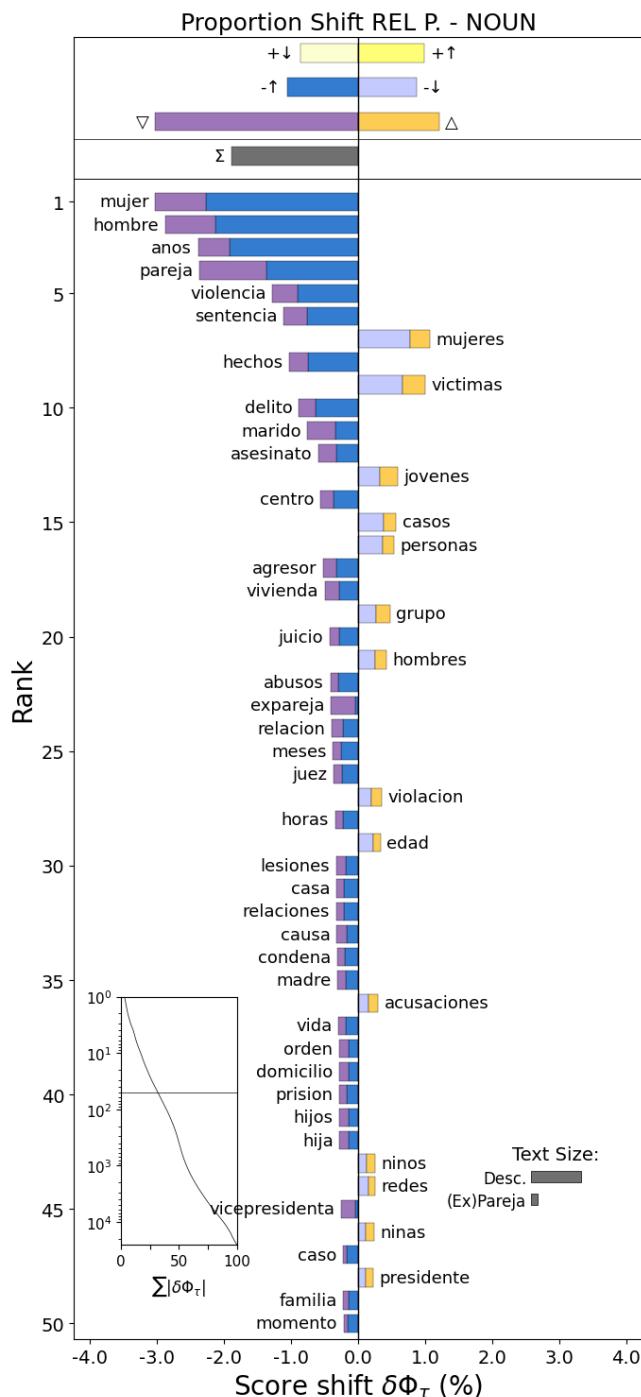
En segon lloc, es mostra la comparació del vocabulari utilitzat en els textos on l'agressor és o ha sigut parella de la víctima (part esquerra de tots dos gràfics) en contrast amb els textos on l'agressor és un desconegut (part dreta dels gràfics), segons la categoria gramatical.



Comparació entre el vocabulari emprat en textos d'agressions on l'autor és o ha sigut parella de la víctima o un desconegut en funció dels **adjectius** utilitzats



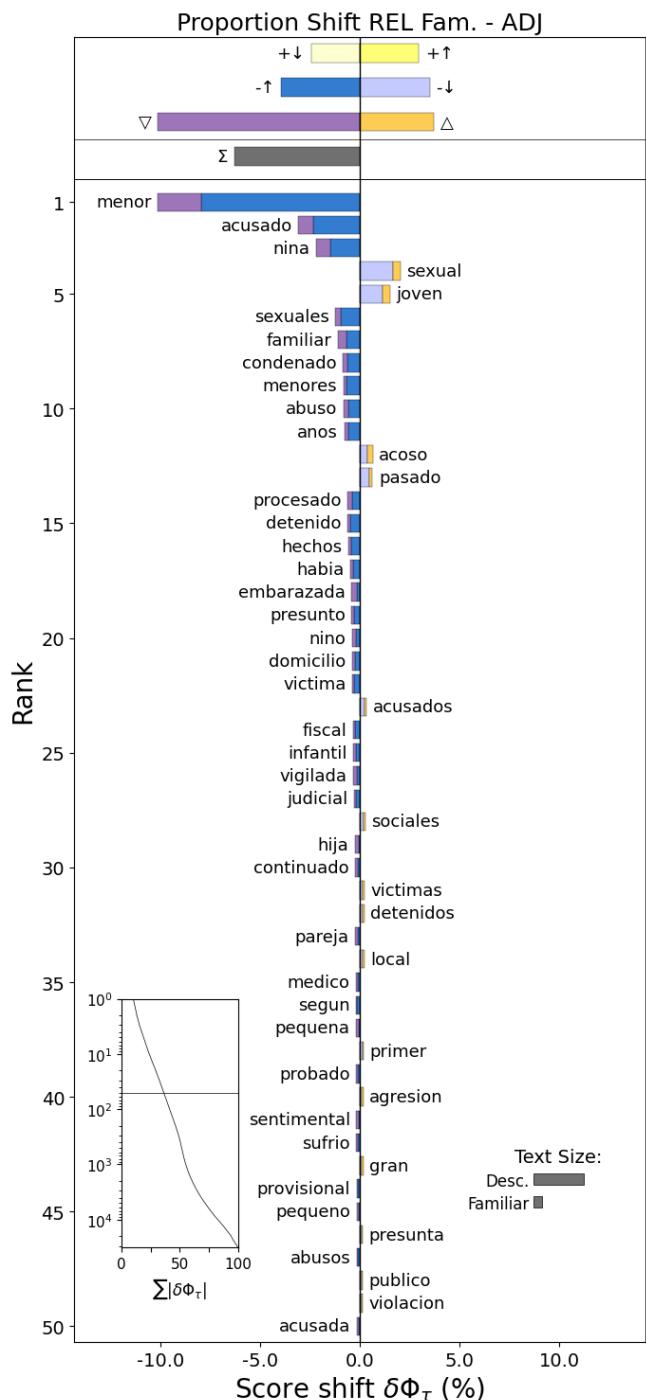
Comparació entre el vocabulari emprat en textos d'agressions on l'autor és o ha sigut parella de la víctima o un desconegut en funció dels **adverbis** utilitzats



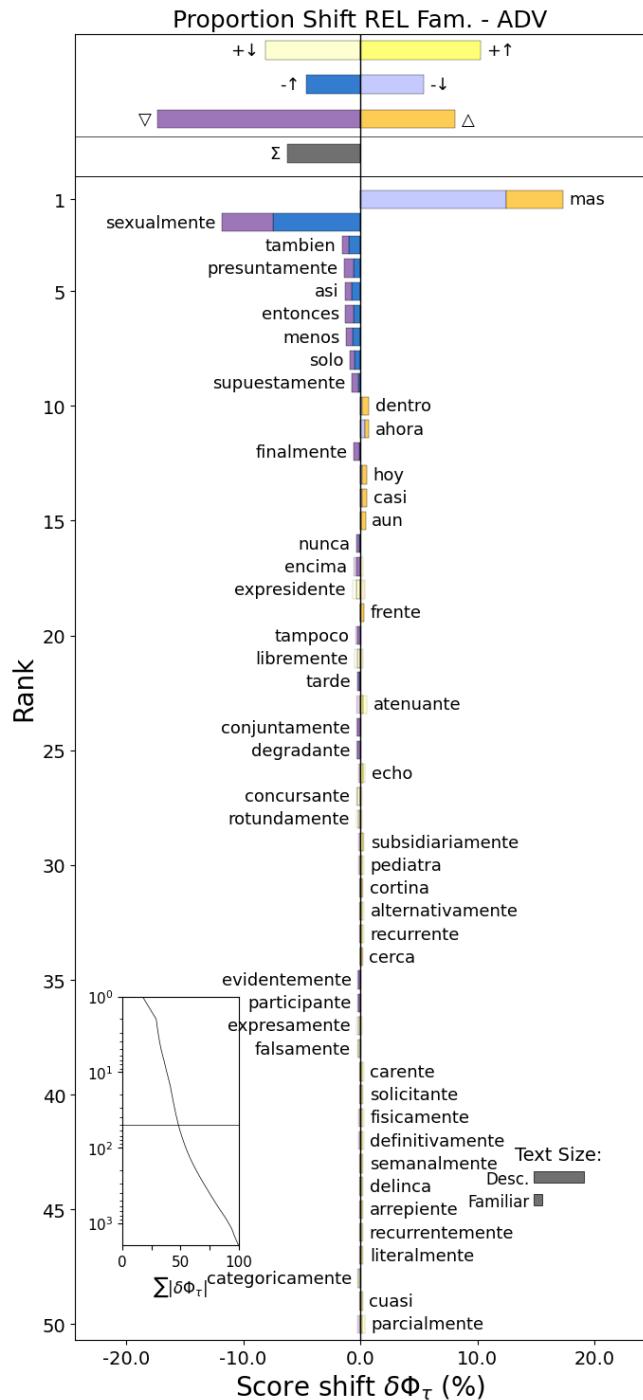
Comparació entre el vocabulari emprat en textos d'agressions on l'autor és o ha sigut parella de la víctima o un desconegut en funció dels **noms** utilitzats

Comparació entre el vocabulari emprat en textos d'agressions on l'autor és o ha sigut parella de la víctima o un desconegut en funció dels **verb**s utilitzats

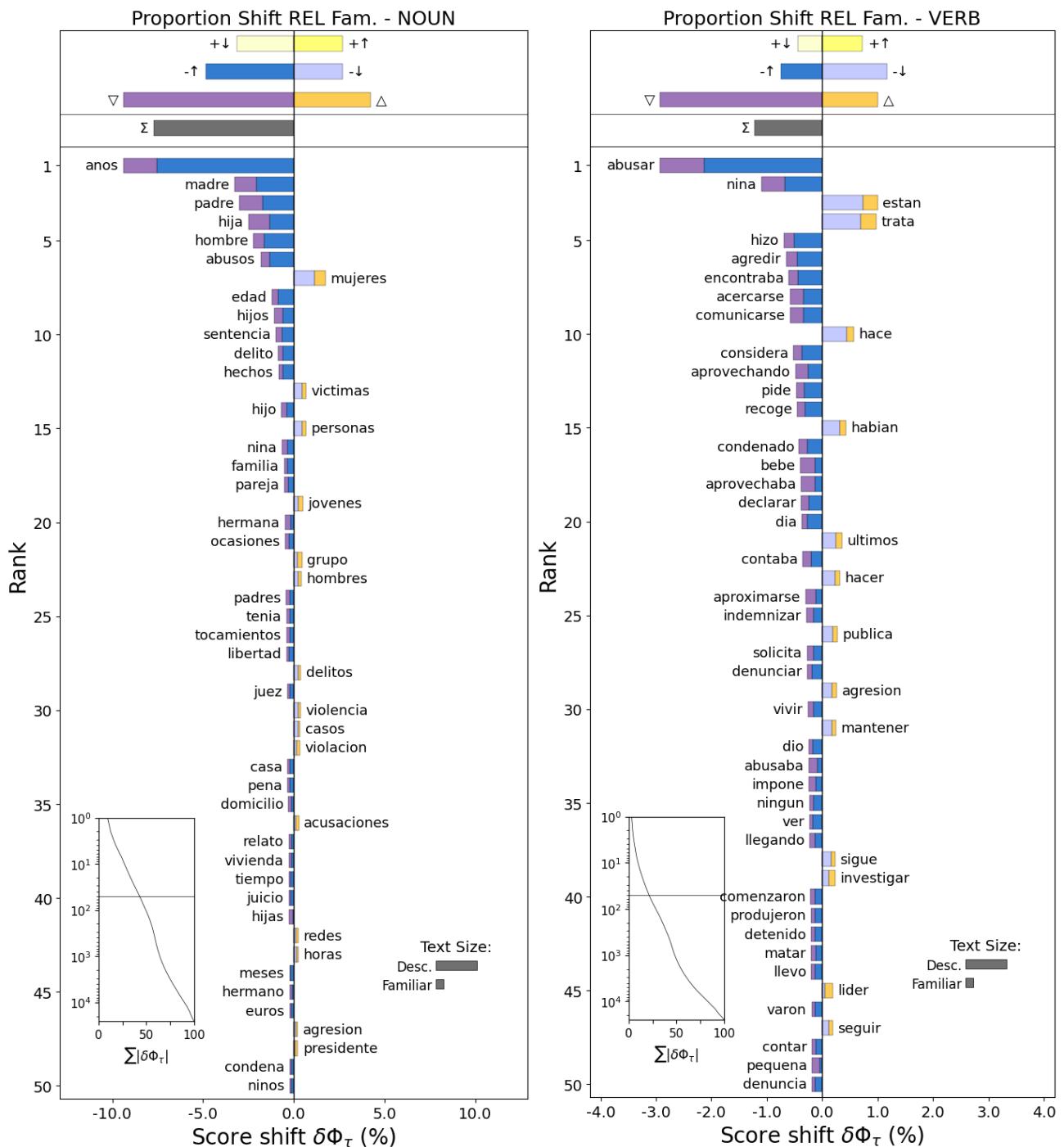
I finalment, es mostra la comparació visual del vocabulari emprat en els textos on l'agressor és un familiar (part esquerra de tots dos gràfics) en contrast amb els textos on l'agressor és un desconegut (part dreta dels gràfics), segons la categoria gramatical.



*Comparació entre el vocabulari emprat en textos d'agressions on l'autor és un familiar de la víctima o un desconegut en funció dels **adjectius** utilitzats*



*Comparació entre el vocabulari emprat en textos d'agressions on l'autor és un familiar de la víctima o un desconegut en funció dels **adverbis** utilitzats*



Comparació entre el vocabulari emprat en textos d'agressions on l'autor és un familiar de la víctima o un desconegut en funció dels **noms** utilitzats

Comparació entre el vocabulari emprat en textos d'agressions on l'autor és un familiar de la víctima o un desconegut en funció dels **verbos** utilitzats