Analysis of online news articles: the coverage of sexual violence in Spanish Media

Budan Budan, Marilena

Curs 2020-2021

Director: CARLOS ALBERTO ALEJANDRO CASTILLO OCARANZA

MATHEMATICAL ENGINEERING IN DATA SCIENCE

upf. Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Treball de Fi de Grau

Analysis of online news articles: the coverage of sexual violence in Spanish Media

TREBALL FI DE GRAU DE

Marilena Budan Budan

Director: Carlos Castillo

Mathematical Engineering in Data Science

Curs 2020-2021

upf. Universitat Pompeu Fabra Barcelona

Escola d'Enginyeria

Acknowledgements

First and foremost, I would like to express my sincere gratitude to the supervisor of this project, Carlos Castillo. This dissertation could not have been completed without your constant guidance and support, and for this, I thank you.

I would also like to thank my friend and colleague Berta, and my cousin Radu for their advice and moral support when dealing with this emotionally draining journey.

To my family, I want to sincerely thank you for supporting me throughout my entire academic journey.

Most importantly, none of this could have happened without the contribution of all the researchers that based the present project; thank you all for your work.

Abstract

More than 10% of women in Spain have suffered from sexual violence instances at least once in their lives. Nevertheless, there are huge prejudices and stigma in society regarding this type of offences, which influence victim's decision when it comes to reporting the crime. The present study analyses news articles, from the most popular media outlets in Spain, to get insights on the coverage of sexual violence, with a focus on the perpetration of harmful preconceptions and myths surrounding sexual violence by means of supervised classification and Natural Language Processing techniques. The study finds that media outlets over-represent extreme cases of sexual violence, meaning those that satisfy the preconceptions' requirements, to maximize their profit instead of providing a fair portrayal of the reality, keeping stigmas alive. This paper will look into sexual violence coverage, by first creating a dataset with news articles related to the sexual violent crimes, then putting them into clusters according to the cases they represent; the cases' main characteristics are then extracted in order to finally compare them with official statistics and social prejudices.

Resum

Més del 10% de dones residents a Espanya han estat víctimes de violència sexual al menys un cop a la seva vida. No obstant, la societat actual segueix mantenint molts prejudicis i estigmes sobre aquest tipus de crims; les mateixes aprensions tenen una gran influencia sobre les víctimes a l'hora de decidir si denunciar els fets. Aquest projecte analitza articles de violència sexual, publicats pels diaris més populars d'Espanya, per entendre la representació dels casos de violència sexual als mitjans de comunicació; l'estudi posa èmfasi en la perpetuació d'estereotips i mites mitjançant l'ús de tècniques d'aprenentatge profund i processament de llenguatge natural. L'estudi ratifica com els mitjans de comunicació destinen una major part dels articles a casos extrems de violència sexual, els quals coincideixen amb els prejudicis socials, ja que incrementen els seus ingressos. L'anàlisi de la representació de la violència sexual s'ha dut a terme començant per la recol·lecció d'un conjunt d'articles sobre la qüestió estudiada, seguit de l'agrupació de les notícies cobrint el mateix esdeveniment, i finalment, s'han extret les seves característiques principals per comparar-les amb estadístiques oficials i prejudicis socials.

Resumen

Más del 10% de mujeres residentes en España han sido victimas de violencia sexual al menos una vez a lo largo de sus vidas. Sin embargo, la sociedad actual sigue conservando muchos prejuicios y estigmas sobre este tipo de crímenes; dichas aprensiones tienen una gran influencia sobre las víctimas en el momento de decidir si denunciar los hechos. Este trabajo analiza artículos sobre violencia sexual, publicados por los diarios más populares de España, para entender la representación de los casos de violencia sexual en los medios de comunicación. El estudio pone énfasis en la perpetuación de estereotipos y mitos mediante el uso de técnicas de aprendizaje profundo y procesamiento del lenguaje natural. El estudio ratifica como los medios de comunicación destinan una mayor parte de sus artículos a casos extremos de violencia sexual, los cuales coinciden con los prejuicios sociales, puesto que incrementan sus ingresos. El análisis de la representación de la violencia sexual se ha llevado a cabo empezando por la recolección de un conjunto de artículos sobre la cuestión estudiada, seguido de la agrupación de las noticias que cubren el mismo suceso, y finalmente, se han extraído sus características principales con tal de compararlas con estadísticas oficiales y prejuicios sociales.

Table of contents

List of figures

List of tables

1 INTRODUCTION

1.1 Background

News articles play an important role in today's society; people consume them on a daily basis and the way the information is described can affect how we see and understand the nature of reality itself. There have been many longitudinal studies demonstrating that media outlets and news articles can manipulate and have an impact on our beliefs (Fitzpatrick, 2018; Morgan, 2018; Stanford History Education Group et al., 2016) especially those concerning sensitive topics such as sexual violence.

The World Health Organization defines sexual violence as "any sexual act, attempt to obtain a sexual act, unwanted sexual comments or advances, or acts to traffic, or otherwise directed, against a person's sexuality using coercion, by any person regardless of their relationship to the victim, in any setting, including but not limited to home and work" (Daher, 2003). Over the past decade there has been a dramatic increase in the number of known crimes against freedom and sexual indemnity in Spain. According to the Spanish Interior Ministry, in 2019 these types of crimes increased by 11.16% with respect to the previous year, reaching a total of 15,319 cases (Ministerio de Interior de España, 2019). These statistics should not be understood only as mere numbers, rather, they should emphasize those who suffered and the steps that can be taken to comprehend and solve these criminal behaviors.

The conflict arises on account of the magnitude of events that happen worldwide every day. Media outlets are forced to decide which events are worthy of being published; therefore, they determine the way reality is presented to the society, which may end up with a biased representation of the actual facts. As articles help shape public opinion (Damstra, 2019), sexual violence representation in the media is a relevant topic for our society.

1.2 Problem statement

The present study arises from the concern on the impact of sexual violence's coverage and representation in the Spanish media. How are these numerous crimes mapped into media outlets' publications? Do they reveal the real incidents and offences that happen on a daily basis or only some 'privileged' cases are widely broadcasted? Do articles feedback on the stereotypes and myths amplifying the preconceptions about sexual violence victims and perpetrators?

Answering these questions can provide clarity and transparency in this, mostly, taboo topic; making it easier to identify behaviors that go against freedom and sexual indemnity, generating more empathy with the victims and helping them to step forward. Therefore, we believe that media outlets should bring awareness towards this field by ensuring a fair representation of reality.

Since fairness is a subjective concept, studies of this field (Aroustamian, 2020; Evans, 2018; O'Hara, 2012) are mostly based on the manual inference of articles' characteristics, due to redaction's subjectivity, requiring time-consuming tasks seeking to automatize them.

The present study was focused on the comparison of coverage of cases by media outlets and the obtention of insights related to the writing style of the authors. Trying to detect whether there are some types of cases more popular than others—e.g., do harassment cases receive the same attention as assault cases?—and whether journalists tend to manifest doubt about information expressed in the articles or to understate it by means of expressions, such as euphemisms.

1.3 Objectives

The objectives of this study are dived into three main categories:

A) Dataset creation

The first objective consists in the creation of a dataset of news articles deemed to describe sexual violence cases published by the most read online media outlets in Spain.

B) Clustering of news articles

The second aspiration is to create an automatized model able to identify and group articles that are about the same case.

C) Articles and cases analysis

The last goal is to analyze the coverage and writing styles of the articles and cases of the dataset. To study the coverage of sexual violence cases, we will analyze if media outlets give more attention to certain types of cases by comparing the representation found in our dataset with official statistics; whereas to examine the writing styles of these articles, we will be focusing on the information provided in the articles and the presence of popular stigmas, expression manifesting uncertainty, and euphemisms in the data collection.

Both studies will provide us insights on the general characteristics of the sexual violence articles, aiming to answer the following questions:

- Do media outlets offer the same coverage to all types of sexual violence? Are all types of bonds between the victim and the perpetrator represented accordingly to the reality? Do sexual violence crimes receive more attention when according to the type of place they occur? If there are biases in the representation of sexual violence cases, which are the characteristics more mediatized?
- Which information is typically provided in sexual violence related articles? Do journalists include age and nationality, or the time and location in which the crime took place? If they do, in which parts of the article?
- Can we find a generalized behavior of including expression related to euphemisms and doubt in these articles? Do they tend to contain stigmas about sexual violence demonizing the perpetrator or emphasizing the possible vulnerabilities of the victim?

1.4 Structure of the present document

The present document is divided into 6 chapters briefly explained below:

- Chapter *1 INTRODUCTION* describes the background problem and sets the goals pursued in the present project.
- Chapter *2 RELATED WORK* contains a review on the main information considered and the literature studied.
- Chapter *3 METHODOLOGY* details the process followed for the development of the project and the methods used in each of its stages.
- Chapter *4 RESULTS* presents the results obtained by following the methodology presented in the third chapter.
- Chapter *5 DISCUSSION* relates the results previously presented with the literature reviewed in the second chapter.
- Chapter *6 CONCLUSIONS & FUTURE WORK* ends with the global conclusions about the goals set at the beginning of the project and proposes further improvements and future work.

2 RELATED WORK

2.1 Sexual violence statistics in Spain

Sexual violence offences are a public concern as more than 2 million women in Spain, 16 or older, have suffered from sexual violence crimes at least once in their lives (Ministerio de Igualdad de España, 2020). In recent years, the model of sexual violence crimes has been compared with an iceberg (Andres-Pueyo et al., 2020; López, 2017), its tip representing the observable events, while the hidden part emphasizing the magnitude of the uncertainty in this field. It is estimated that only 13.7% of Spanish women who have suffered from sexual assault have reported the crime (Ministerio de Igualdad de España, 2020).

The National Institute of Statistics (INE) offers information about the number of incarcerated people each year due to sexual violence crimes classified by types of offences. There are three main categories of sexual violence defined as:

- Sexual harassment: Unwelcome sexual advances, requests for sexual favors, and other verbal or physical conducts of sexual nature. Harassment does not involve penetration.
- Sexual assault: An act of physical, psychological, and emotional violation in the form of a sexual act, inflicted on someone without their consent. It can involve the forcing or manipulation of someone to witness or participate in any sexual acts.
- Sexual abuse: An act of violence inflicted by the attacker against someone they perceive as weaker than them. It is a crime committed deliberately with the goal of controlling and humiliating the victim.

Table 1: Criminals incarcerated due to sexual violence crimes during 2019 disaggregated by type of offence

Type of sexual violence	People incarcerated	Proportion
Assault and rape	488	20.35%
Abuse	1383	57.67%
Harassment	527	21.98%
TOTAL	2398	100%

The most important statistical study concerning sexual violence in Spain is *Macroencuesta de Violencia contra la Mujer, 2019* (Ministerio de Igualdad de España, 2020); an anonymous survey conducted to 9,568 women, performed every four years since 1999, willing to discover the proportion of women, 16 or older residing in Spain, that are suffering or have suffered sexual violence. Among the information reported, results present information about the type of

bond between the victim and the perpetrator—summarized in the subsection *Victim-perpetrator bond* of part *3.4.1*—along with data about the places where these crimes took place when the victim and the perpetrator did not had a relationship prior or at the time of the sexual violence offence—information presented in *Table 3*.

Table 2: Prevalence of different type of bonds between the victim and the perpetrator of sexual violence

Victim-perpetrator bond	Prevalence in Macroencuesta 2019
Relationship	72.20%
Relative	6.00%
Friend or acquaintance	14.03%
Stranger	10.86%
Unclassified	0.05%

Table 3: Prevalence of types of places where sexual violence occurs when the victim and the perpetrator did not have a relationship-type of bond

Type of place	Prevalence in Macroencuesta 2019
Public	36.3 %
House	65.4 %
Workplace	7.1 %
Educational	6.2 %
Leisure	15.6 %
Unclassified	0.3 %

2.2 Representation of sexual violence in the media

As awareness about sexual violence has been on the rise, it is increasingly recognized as a serious, global public health concern, since crimes against freedom and sexual indemnity have serious consequences on victim's mental health. Consequently, many studies analyze the social connotations attached to sexually violent offences (Flanders et al., 2019; Murray et al., 2016).

Due to the essential function that news articles play in modern culture, media outlets are able to manipulate and bias readers' ideas and beliefs (Fitzpatrick, 2018; Morgan, 2018). Therefore, sexual violence's representation in news articles is a key factor for understanding the spread of prejudice, myths and stereotypes surrounding this topic (De Benedictis et al., 2019; DiBennardo, 2018; Evans, 2018; O'Hara, 2012; Walton, 2020) and the connotation of the language used in the description of cases involving sexual violence offences (Aroustamian, 2020; Conboy, 2007). Some of the main findings of the former research are the following ones:

- Media outlets generally attach monstruous imagery along with the description of sexual predators, demonizing and labelling perpetrators as people with mental illness or alcohol problems perpetrating the following misconceptions (O'Hara, 2012; Walton, 2020):
    - People that carry out sexually violent crimes suffer mental health problems; therefore, these conducts should not be generalized.
    - Since sexual violence offence are performed by 'mad' people, they are considered as another crime which is rarely treated with the attention it requires.
- Victims are generally idealized according to both sexual violence stereotypes, and social implicit bias towards white privileges (DiBennardo, 2018):
    - There is a considerable bias towards the spread of sexual violence cases involving celebrities and focusing on the film industry and fashion sector (De Benedictis et al., 2019).
    - There's a significant favoritism towards white women's experiences and a major presence of Caucasian demographics in the analysis of attributes of victims represented in news articles (De Benedictis et al., 2019; Evans, 2018).
    - Victims are classified into one of these polarized categories: either idealized virgins or promiscuous women who do not deserve to be treated as victims; the latter are also sometimes characterized as having demanded the attention of their attacker (O'Hara, 2012).
- There are different prejudices along each type of sexual violence.
    - Sexual assault and rape victims are being extremely victimized by the media by "being blame-worthy for the pain inflicted on them by their attackers" and represented as promiscuous women (Aroustamian, 2020; O'Hara, 2012).
    - Sexual harassment stereotypes perpetuated by the media involve the understatement of experiences and the normalization of these types of crimes by downsizing their gravity and being referenced just as "feared situations" (Walton, 2020).
    - Preconceptions around sexual abuse cases depend on the scenario: when abuses occur in leisure spaces, such as a bar or a nightclub, most of the times, news articles blame victims while excuse perpetrators; on the contrary, the

representation of child abuse cases in news articles is focused on the "predatory nature of offenders" (DiBennardo, 2018), usually described as violent crimes "committed by strangers and poor men of colour" (Aroustamian, 2020).

- The language used in sexual violence-related articles perpetuates myths and popular misconceptions regarding these types of crimes (O'Hara, 2012) by presenting a high presence of euphemisms and confusing language (Aroustamian, 2020) such as "stealing someone's virginity". These idioms and expressions shift away the attention from the sexual violence case to the emphasis of stereotypes.

2.3 Twitter analysis

Twitter is an online free platform used for microblogging and social networking. It was funded and launched in 2006; fifteen years later it reached 186 million users. Twitter defines itself as "Twitter is what's happening and what people are talking about right now."

As a consequence of its popularity, Twitter affects many aspects of the society such as interpersonal communication, news consumption, and journalism (Fitzpatrick, 2018); Twitter's popularity places it in the spotlight of interdisciplinary fields of study such as influence in social polarization (Walter et al., 2020); the potential use for the detection and prevention of suicide (Leiva & Freire, 2017); and the use of media manipulation in politics (Ouyang & Waterman, 2020) among others.

This platform is not only used as a social network, but also as a source of information (Kwak et al., 2010) considering that media outlets use Twitter for redirecting users to their official webpages (Calvo et al., 2020); multiple studies remark that people switched their attention from traditional media forms to social media platforms for the consumption of news in recent years (Fitzpatrick, 2018; Gottfried & Shearer, 2016).

2.3.1 Twitter as a source of data

The information posted on Twitter can be accessed through three different levels:

- The public level: As it is a free platform, there are different frameworks designed for the extraction of information that can be consumed from the public level; this option typically outputs few attributes about each publication.
- The private level: The platform offers a private level to their Application Program Interface (API) designed for researchers, developers and enterprises; in terms of data

extraction, this level provides more information than the public one. However, it requires having access to a Twitter Developer Account and the plan selected determines the limitations with respect to data retrieval.

- Web parsing: The third level is not considered nor provided by the platform; it consists in the acquiring of information by parsing of the content of their webpage.

2.4 News articles analysis

According to the Reuters Institute (2020), 83% of news articles' consumers in Spain prefer online sources, and 63% of them are concerned about the veracity of the information broadcasted (Newman et al., 2020). Previous studies have evidenced the impact of news articles in societies' beliefs (Fitzpatrick, 2018; Morgan, 2018; Stanford History Education Group et al., 2016) motivating many researchers to investigate this topic.

The main approach followed for the automated analysis of news articles involves Natural Language Processing (NLP), which is a subfield of Artificial Intelligence that combines Linguistics and Computer Science for the analysis and comprehension of texts written in human languages (natural language).

NLP can be applied to a broad range of areas including machine translation, text classification, spam detection, extraction of information, and summarization; However, one of the most popular topics in articles analysis in the area of NLP is the detection of fake news (Morgan, 2018; Smitha & Bharath, 2020; Sriram, 2020) and the study of the coverage and content analysis of news about specific topics, such as the COVID-19 pandemic (Basch et al., 2020; Y. Chen et al., 2013; Hart et al., 2020).

News articles analysis require a prior transformation of the text, since Machine Learning models do not take natural language texts as input. Some of the most used text representation techniques are:

- One-hot encoding and Count vectorizer: these techniques represent a collection of documents as a matrix, indicating whether each term of the entire collection appears in each document (one-hot encoding) or its frequency in a given document (count-vectorizer). These techniques are mainly applied in supervised tasks and keywords analysis (Smitha & Bharath, 2020).
- Bag-of-N-grams: this technique takes into account the order of the terms in documents; it defines a minimum unit, named gram, and represents a text as all the N consecutive

grams that it contains. E.g., the sentence ‘I love natural language processing’ would be represented as [‘I love’, ‘love natural’, ‘natural language’, ‘language processing’] considering bag-of-2-grams.

- Named entities recognition: technique that represents a document as the set of named entities that it contains (Gupta, 2011).
- TF-IDF weighting: this technique assigns a weight (representing its relevance) to each term of a vector considering the frequency of the term in both the entire collection and in the given document. TF-IDF weighting’s is applied mostly for document comparison tasks (Qaiser & Ali, 2018).
- Term embedding: these techniques learn how to represent terms as vectors in a feature space such that the terms that are closer in the vector space have similar semantic and syntactic information. Term embeddings are mainly used for classification and the analysis tasks (Sriram, 2020).

Having obtained the vectorial representation of data collection, different types of analysis can be applied depending on the goal of the study, some common scenarios and techniques are:

- Coverage analysis: a common method is to filter documents by keywords occurrences (H. Chen et al., 2020; Hart et al., 2020).
- Supervised articles classification: most popular solutions involve the application of Machine Learning models that learn the classification given the ground truth (Smitha & Bharath, 2020); a popular example of binary classification is detection of spam emails (Kumar et al., 2020).
- Extraction of main topics: unsupervised clustering models are typically applied for extracting the main groups intrinsic in the dataset, in the case of news articles, these tend to be the topics of discussion (Huang, 2008).
- Documents comparison and detection of near-duplicate texts: computation of similarity or distance metrics for approximating the similarity between a pair of documents for tasks like plagiarism detection (Malik et al., 2020).

Collectively, these NLP techniques outline the increasing popular interest in text processing and news analysis.

3 METHODOLOGY

3.1 Pipeline

The methodology followed for achieving the goals stated in section *1.3 Objectives* is summarized in *Figure 1: Methodology pipeline*.

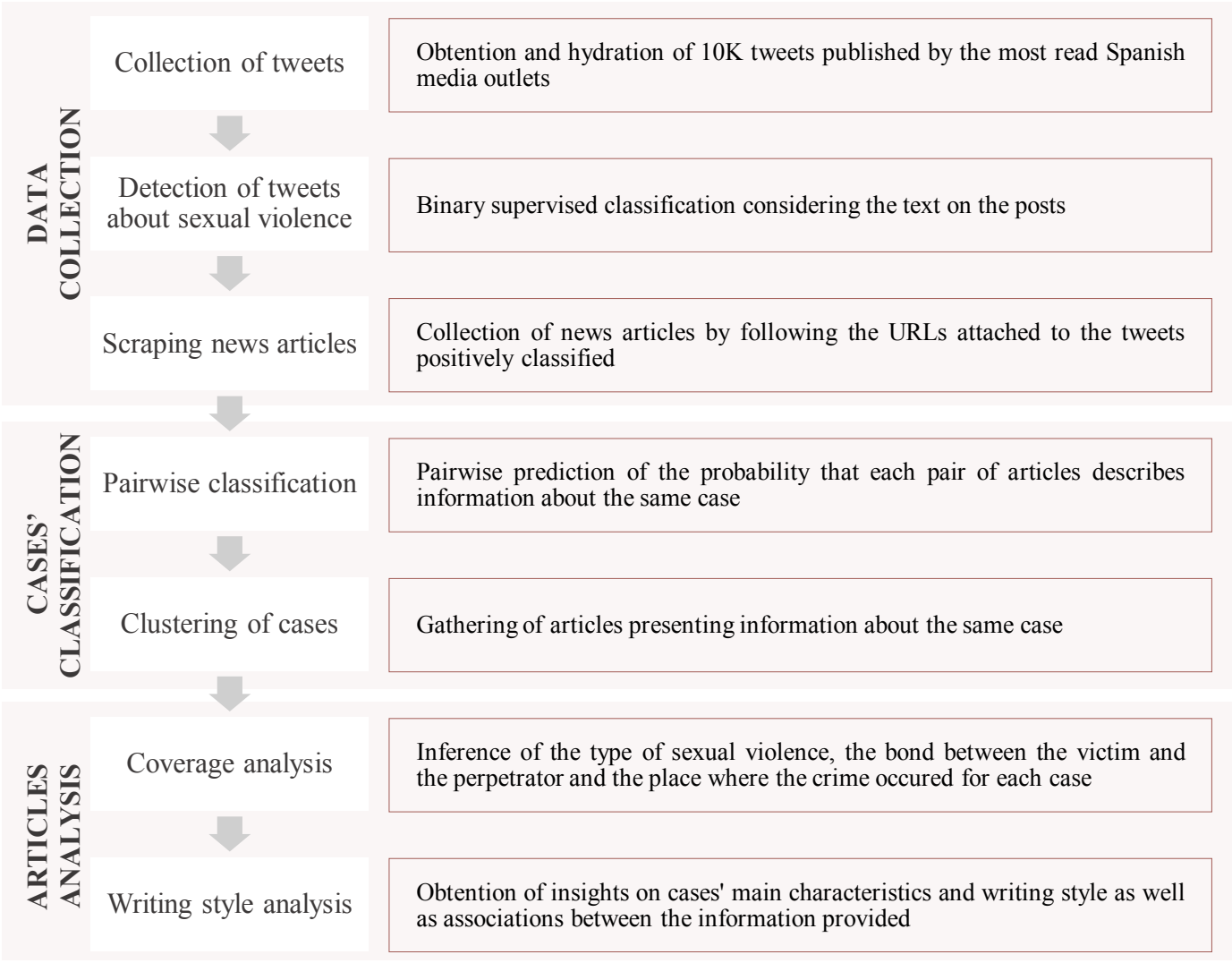

Figure 1: Methodology pipeline

The methodology can be split into three main steps: (A) data collection, the creation of a dataset with news articles deemed to describe sexual violence, (B) articles' classification into cases, and (C), articles' analysis focusing on the coverage of cases and the obtention of insights about the most common characteristics and writing styles.

3.2 Data Collection

The present work is based on the analysis of news articles published by 15 of the top 16 most read online news media outlets in Spain, according to the Reuters Institute Digital News Report *Figure 2* (Newman et al., 2020). The report highlights the 16 most popular news outlets. Due to its focus on sports-related news articles, the Marca newspaper was not considered, as its relevance was deemed insufficient for the purposes of this research. Therefore, our analysis will be centered on the publications of the following media organizations: El País, El Mundo, 20 Minutos, El Diario, Antena 3, El Confidencial, OKDiario, La Vanguardia, La Sexta, Telecinco, RTVE, ABC, El Periódico, Cadena SER and La Razón.

Figure 2: Weekly reach offline and online of the top brands in Spain (Newman et al., 2020)

The pipeline shown below in *Figure 3* describes the process conducted to obtain our dataset.

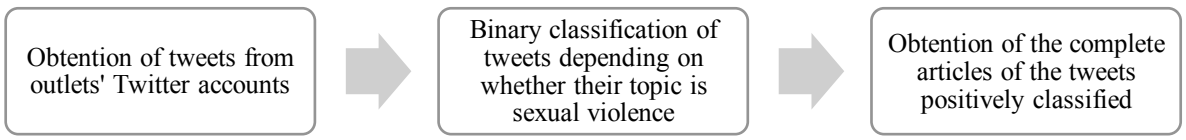

Figure 3: Data collection creation pipeline

Media outlets will use social networks, such as Twitter and Facebook, in order to broadcast their content with the goal of increasing the traffic of their websites (Ahmad, 2010).

Twitter has played an essential role in the dataset creation process; the articles used in the project have been chosen from the individual tweets published by the official accounts of the previously mentioned media outlets.

A total of 10,000 tweets have been collected from each account. Subsequently, the tweets were binary classified, using a supervised model, to describe their relation to sexual violence: Those labeled as 1 were deemed to describe sexual violence cases, and those labeled as 0 were deemed not to describe such violence. Finally, only the positively classified tweets were used to collect

the news articles from each news outlet's webpage by following the URL attached to each tweet.

3.2.1 Twitter scraping

The first step we took in the data collection phase was the selection of relevant Twitter accounts for each outlet we considered. Most media outlets have multiple accounts on Twitter and each account is dedicated either to a different type of content—e.g., some of the official Twitter accounts of Antena3 are @antena3com, @A3Noticias, @Antena3Deportes, @LaVozAntena3—or to a territory they operate in—e.g., for El País there are @el_pais, @elpais_america, @elpaiscatalunya, among others. In both cases, the account selected has been the one dedicated to news articles with the most global information focusing on Spain.

Table 4: Media outlets selected, their corresponding Twitter account, and account's description extracted on 2021-05-10

MEDIA OUTLET	ACCOUNT	DESCRIPTION
El País	@el_pais	La mejor información en español. Con nuestra mirada puesta en España, Europa y América.
El Mundo	@elmundoes	Cuenta oficial de EL MUNDO
20 Minutos	@20m	Cuenta oficial de 20minutos, el medio social y ciudadano. Información, análisis y contacto personal con los lectores las 24 horas
El Diario	@eldiario	#TeExplicamosLaNoticia Medio que hace periodismo con carácter y sin censura ¡Únete a la conversación!
Antena 3 online	@A3Noticias	Toda la información a un click, en http://antena3noticias.com La actualidad, en tu móvil, a través de http://t.me/A3Noticias
El Confidencial	@elconfidencial	Únete a los lectores influyentes. Suscríbete a El Confidencial: http://elconfidencial.com/suscribete/ #únetealconfi
OK Diario	@okdiario	Vienen a por nosotros ¡Apóyanos! https://inconformistas.okdiario.com
La Vanguardia	@LaVanguardia	Pluralidad, rigor y calidad desde 1881. La Vanguardia es el lugar donde todos y todas nos encontramos.
LaSexta	@sextaNoticias	El twitter de laSexta | Noticias. Te contamos todo lo que ocurre en el momento que ocurre.
Telecinco	@informativost5	Perfil oficial de Informativos Telecinco |
RTVE News	@rtve	La actualidad al minuto en @rtve
ABC	@abc_es	Diario ABC.
El Periódico	@elperiodico	Entender +
Cadena SER	@La_SER	Cadena SER - La radio
La Razón	@larazon_es	Información / Innovación / Emoción Cuenta oficial del diario LA RAZÓN. RT y HT no significan necesariamente acuerdo.

Having determined the most suitable account for the current goal, a total of 10,000 tweets were collected, on the date 2020-09-16, from each Twitter media account using a python script—refer to *ANNEX 1. Data and code release* for further information. Collecting the same number

of Twitter posts for each outlet implies differences in time ranges since the collection is dependent on the frequency of the tweets each account posts.

There are multiple options for the extractions of information from Twitter—as stated in section *2.3.1 Twitter as a source of data*. Due to restrictions of the private level (it limits the collection of tweets to 3,200 from each user's timeline) data was collected by means of the public level, which provides general information about each tweet—i.e., tweet_id, the link to access that tweet, the username of the account, the text of the tweet, the date it was posted, the number of retweets and favorites of the tweet, the mentions and the hashtags that the text contains. A total of 100,000 tweets were collected from each user's timeline.

3.2.2 Tweets hydration

The data, collected using the public level of access, does not include the URLs of the attached articles. Therefore, to obtain the complete information available about each tweet, they have been hydrated[1] using the private level.

The hydration process outputs, for each file containing the summary of 1,000 tweets, a JavaScript Object Notation (JSON) file with one tweet per line—jsonlines—containing the extended information collected from the Twitter API.

Tweets' hydration phase is prone to reducing the data size because some of the tweets available during the scraping may have been erased before the hydration takes place.

3.2.3 Tweets classification

The main topic of news articles referenced in tweets can be obtained from the tweet's text (Calvo et al., 2020). Since we are interested in news articles about sexual violence, classifying the tweets is the optimal solution.

The scope of this section is the classification of each tweet collected in the *3.2 Data Collection* phase in order to distinguish those tweets which present information about sexual violence cases by virtue of the associated text within the individual tweets.

[1] Technique that consists in obtaining the complete details of a tweet by querying it to Twitter API using the tweet id as an identifier.

Data transformation

Textual data has to be transformed before being fed to Machine Learning models. Normalization of the tweets was performed, before the data transformation, by lowering all the characters, by removing symbols, punctuations, and stopwords[2]; reducing the number of unique terms in the data collection and the computational cost of the whole process.

Normalized tweets were then transformed into numerical vectors using a count vectorizer, which creates a sparse matrix of size NxM, N being the number of records and M the number of unique terms in the dataset. Each position $[i, j]$ of the matrix specifies the number of times term j appears in record i. Remarkably, this simple transformation has outperformed more sophisticated ones, such as the binary classification of fake news (Smitha & Bharath, 2020).

Classification approach

Machine learning approaches are broadly divided into supervised and unsupervised learning. Classification models are algorithms that learn to classify information in different classes; supervised classification models learn to distinguish classes given the labels that determine the ground truth of each observation, while unsupervised learning models do not take labels as input for learning and, the model infers the class by discovering patterns and intrinsic properties of the data (Reddy et al., 2018).

The diversity of the posts published on Twitter by news media outlets (Calvo et al., 2020) implies having a dataset with a vast amount of topics, making it challenging to attain a precise classification by unsupervised means. At the same time, the dimension of the dataset (150,000 records) requires a considerable labeling effort. For this reason, the approach followed uses a classification model for fastening the labeling part until obtaining the desired classification performance.

Model selection

The model selected for this task is a binary logistic regression—shown by Samitha & Bharath (2020) to provide good results for detecting fake news (text classification). Binary logistic regression is a statistical model that predicts the probability of belonging to a class given independent variables by applying the sigmoid function *(1)* to a linear combination of the input variables (Harrell, 2015).

---

[2] Frequently used words that do not provide useful information for the computational purpose, i.e., 'and', 'or', 'the', among others.

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

The output of the sigmoid function is an approximation of the probability of belonging to the class labeled as 1. To obtain a classification out of this probability, a threshold—typically 0.5—is selected; the dichotomous output is obtained by assigning label 1 to those probabilities that are equal or greater than the threshold, while those that are lower are classified as 0 (Harrell, 2015).

Data labeling

For training the supervised binary logistic regression model, each tweet of the training set has to be associated with a label specifying whether sexual violence is the topic of that tweet. The codification used for the assignation of labels is shown in the table below.

Table 5: Criteria used for labeling tweets.

Value	Meaning
0	Not about sexual violence
1	About sexual violence
2	Difficult to distinguish if it is about sexual violence or not

Data labeling is a time-consuming task due to the size of the dataset (100,000 rows = 100,000 tweets). Therefore, the classification model was used for speeding up this task by following an iterative approach.

i. Define the training set by manually labeling few rows.

ii. Train the model with the labeled dataset.

iii. Predict the probability—not the class—of the rest of the dataset.

iv. Sort the data in decreasing order according to the probability predicted in the previous step and label some of the top records.

v. Start again from step *ii* adding the new records classified to the training set.

The sorting of the data in the fourth step is needed to improve classification speed since the original dataset contains only a small proportion of records about sexual violence cases.

By repeating this process several times and adding new information for the training, the model improves, providing better outputs and refining the results.

Final classification

Having labeled around 70% of the dataset, we used the model for computing the probability of a Tweet containing information about sexual violence and predicting the class using a threshold = 0.5.

To get the final classification, the results were sorted in descending order of probability, and the top 10% of the tweets were manually checked, providing the definitive classification of tweets.

3.2.4 News articles scraping

The last step of the creation of the dataset represents the collection of news articles from their original sources (websites of news outlets) by following the URLs attached to the tweets positively classified.

News storage format

The format chosen to store the news articles is NewsML-G2, a standardized way of storing articles in an XML format, defined by the International Press Telecommunications Council[3] (IPTC) to provide open standards for the news media.

Table 6: NewsML-G2 xml structure used to store news articles

```
<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">
```

[3] https://iptc.org

```
    <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>
```

As it can be seen in *Table 6* the data stored contains different fields that can be assigned to the following categories:

- Article's identifier: composed from "urn:newsml:[MEDIA]:[EXTRACTION DATE] :[TWEET ID]"
- Extraction and source data: specifying the date of the extraction and the tweet id where the URL of the article was obtained from.
- Article information: canonical URL source, publication DateTime, the source that published the article, the country it has been published in, and article's author.
- Article's heading: the headline and subtitle.
- Article's text: the body of the article.

Articles obtention process

To obtain the news articles in the aforementioned format, the pipeline shown in *Figure 3* has been followed.

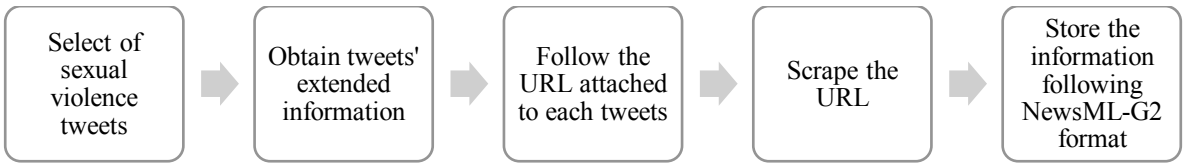

Figure 4: News articles' obtention pipeline

The first step was filtering out tweets not related to sexual violence using the classification obtained from section *3.2.3 Tweets classification*, followed by the obtention of the URLs attached to each tweet from their corresponding JSON file—containing the extended information about tweets.

Having obtained the source link of the articles, each webpage was parsed to extract the contents of the HTML file—since each outlet has its own schema and website structure, the parsing is unique for each news media outlet sampled.

Limitations

Some of the tweets, whose topics are sexual violence, had to be dropped as they either did not include a link at all, or their URL attached pointed to a web page different from a relevant news article—such as their general website or their link for online streaming.

Additionally, some of the tweets presented a 'nofollow' tag to the link they were attaching. The 'nofollow' attribute was invented nearly 15 years ago, and it is used for avoiding any endorsement to the hyperlink being attached—i.e., ranking algorithms do not consider them. Attaching the 'nofollow' attribute to a hyperlink in Twitter makes it untraceable for Twitter's API, even if these links do appear when scraping the HTML of the tweet. To obtain these 'nofollow' URLs, we opened a headless chromium browser (based on selenium), which accessed to tweet's URL and clicked on the link attached in the tweet. Consequently, the chromium browser opened in a new tag the 'nofollow' link attached in the tweet and allowed us to obtain the URL of the news articles. Having obtained the 'nofollow' URLs, that were not available in the JSON files with the extended information of each tweet, news articles can be collected by following the fourth and fifth steps of the obtention pipeline (see *Figure 4: News articles' obtention pipeline).*

3.3 Cases' classification

Since sexual violence cases have a wide range of coverage, it is important to distinguish different individual cases the media is talking about when studying the coverage of sexual violence by the online media. Considering as one case one event, which may be presented in diverse news articles.

A crucial consideration is that our present dataset, made out of news articles about sexual violence cases, contains a similar vocabulary, and it is not a trivial task to cluster those articles that present information about the same case.

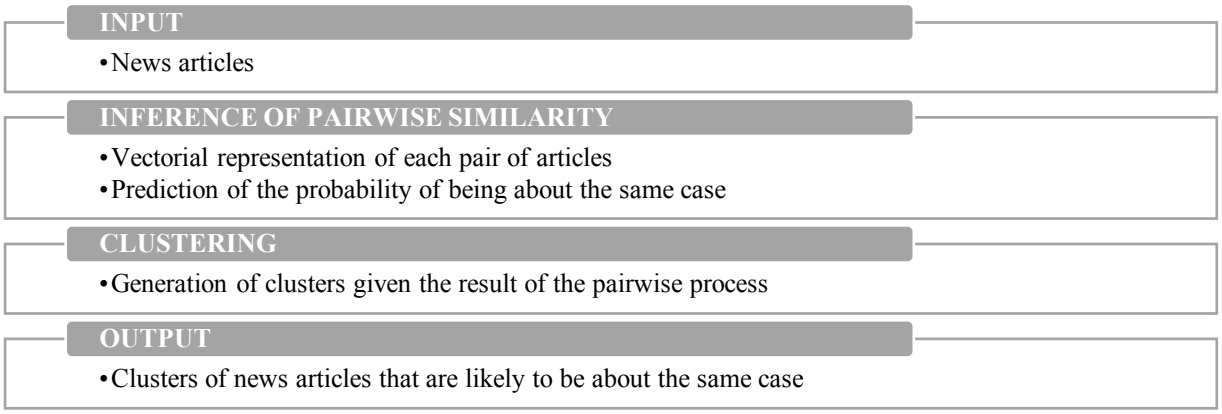

Figure 5: Process for clustering news articles that are about the same case.

The approach followed involves different steps: representing each pair of articles as feature vectors, predicting its probability of being about the same case in a supervised fashion, and clustering all the articles by creating a distance measure considering the previously predicted probability.

3.3.1 Text representation

There are multiple ways of representing text documents, from the bag-of-words representation to document embeddings (Sriram, 2020). For comparing pairs of documents, each article has been represented in various ways to leverage the information that each representation provides.

Named entity recognition (NER)

Named entity recognition (NER) consists of extracting atomic elements present in the text, and their classification into classes such as locations, organizations, and persons, among others. NER has different applications in text processing tasks related to text summarization and classification (Gupta, 2011).

The output of applying NER to a text document is a list of its named entities classified by classes.

Term Frequency Inverse Document Frequency

Term Frequency Inverse Document Frequency (TF-IDF) is a way of representing a text that measures the relative importance of each term in the corpus. TF-IDF perceives the importance of a term as a relative attribute; a term is considered relevant if it does not appear in many documents, computed as the inverse document frequency, and it is important for a given document if it is contained multiple times, considered with the term frequency (Qaiser & Ali, 2018).

TF-IDF cannot be applied to a unique text since it considers all the appearances of each term in the entire dataset. Therefore, the TF-IDF of a given term t and a document d is obtained with the expression *(2)*, where N is the total number of documents in the collection.

$$TF - IDF_{t,d} = tf_{t,d} \cdot idf_t = tf_{t,d} \cdot log\left(\frac{N}{df_t}\right) \quad (2)$$

- $tf_{t,d}$ is the number of times term t appears in document d
- df_t is the number of documents containing the term t

The use of pre-processing techniques that normalize the terms—prior to the application of TF-IDF—is highly recommended for reducing the computational cost and improving texts' representation (Uysal & Gunal, 2014).

Word embeddings

Word embeddings represent words in highly dimensional spaces by capturing both semantic and syntactic features of each word from a corpus. Meaning that two terms will be close in the embedded space if they are somehow semantically and syntactically related (Y. Chen et al., 2013).

- FastText

FastText is a toolkit developed by the Facebook AI Research (FAIR) lab, which uses a deep neural network architecture for learning word embeddings (Joulin et al., 2016).

The model assumes that each term is composed of smaller units, called n-grams, that differ in length depending on the word (Kuyumcu et al., 2019). By taking into account the internal structure of each term, FastText is able to understand the meaning of suffixes and prefixes and to represent terms that were not even present in the dictionary (collection of words of the dataset used for training the model). The model used was pre-trained with the Spanish Unannotated Corpora (Cañete, 2019), which contains more than 3 million words.

- BETO

BERT (which stands for Bidirectional Encoder Representations from Transformers) is a deep learning algorithm, developed by the Google AI Language team, that represents words in an embedded space, focusing on language understanding, and is able to capture syntactic, semantic, and word knowledge (Rogers et al., 2020). BERT consists of two stages: pre-training and fine-tuning. In the pre-training phase, each term is tokenized into word pieces, and three

embedding layers are combined to obtain a fixed-length vector; fine-tunning is used to improve the performance for a given task by adding one output layer (Devlin et al., 2019).

Even if BERT is multi-lingual, the model used for the present work is BETO, a BERT-based language model pre-trained only with the Spanish Unannotated Corpora (Cañete, 2019) that provided better results than the multi-lingual BERT (Cañete et al., 2020).

3.3.2 Similarity metrics

Goodall1 similarity

Goodall is a similarity metric for categorical data that normalizes the similarity of two elements by the probability that this similarity could have happened randomly or by chance, meaning that the coincidence of infrequent attributes in both elements has more importance in their similarity than the match of frequent attributes. Since using this measure is computationally expensive, we can make use of more efficient approaches. Goodall1 is a variant computed as the average of per attribute similarity (Boriah et al., 2008).

Being X and Y two elements and d the number of attributes that each contains, X and Y's similarity is computed as the average of the per-attribute similarity: having a constant weight $w_k = \frac{1}{d}$ and $S_k(X_k, Y_k)$ being the similarity between the k[th] attribute of both elements.

$$S(X,Y) = \sum_{k=1}^{d} w_k \, S_k(X_k, Y_k) \tag{3}$$

$$S_k(X_k, Y_k) = \begin{cases} 1 - \sum_{q \in Q} p_k^2(q) & if \; X_k = Y_k \\ 0 & otherwise \end{cases} \tag{4}$$

The per-attribute similarity is computed if X's and Y's k[th] attributes are equal (4), where $p^2(q) = \frac{f(q) \cdot (f(q)-1)}{N \cdot (N-1)}$, being $f(q)$ the absolute frequency of attribute q (in all elements) and N the total number of elements.

Note that the range of possible values for the per-attribute similarity is $\left[0, 1 - \frac{2}{N\,(N-1)}\right]$, since the minimum per-attribute similarity larger than 0 is $\frac{2}{N\,(N-1)}$, which occurs when an attribute is contained once in only two elements.

Jaccard coefficient

The Jaccard coefficient is a set similarity metric computed as the number of common attributes in both sets, divided by the total number of attributes (Niwattanakul et al., 2013)

$$Jaccard(X,Y) = \frac{|X \cap Y|}{|X \cup Y\,|} \tag{5}$$

The range of Jaccard similarity is $[0,1]$, the extremes are matched when none of the attributes is present in both sets (0) and when all elements appear in both sets (1) without considering their frequencies.

Note that this metric uses a bag-of-words approach since it does not consider attributes' frequency.

Cosine similarity

Given a vectorial representation of two elements, their cosine similarity is the cosine of the angle between both vectors. The metric can be computed from the vectors as their dot product normalized by the module of their cross product (Huang, 2008).

$$CosineSimilarity(\vec{X},\vec{Y}) = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X} \times \vec{Y}|} \tag{6}$$

Word mover's distance

Given the word embeddings of two different texts, the world mover's distance computes the distance between the texts as the minimum cumulative distance that the embeddings of one text have to travel to be transformed into the representation of the second text. Since it is computed using word embeddings, the metric considers the changes that have to be done at a semantic and syntactic level (Bahrdwaj, Saksham; Saxena, 2019).

3.3.3 Features

The classification of cases has been done in a pairwise fashion, meaning that the feature matrix contains a vector (row) for each pair of articles, and it has been used for binary classifying whether those two articles presenting information about the same case. Features were created by combining text representation techniques with similarity metrics—presented in sections *3.3.1 Text representation* and *3.3.2 Similarity metrics*, respectively—following four different approaches.

Set-based similarity features

Documents can be represented by the set of entities they contain. Even if entities do not summarize the entire content of any document, they highlight some of document's most important terms. Therefore, the similarity of two articles can be assessed by applying categorical similarity metrics, i.e., Jaccard coefficient and Goodall1 similarity, between articles' sets of entities.

Before computing the set-based similarity features, named entities sets have been split into two groups: the first one containing the entities collected from the lead part of the article, and the second one containing those named entities found in the body of the article. The lead of a newspaper article tends to summarize the main content of an article, while in the rest of the body, the same information is typically explained in an extended way (Conboy, 2007). For this reason, the entities found in the lead may be more relevant than the other ones and each similarity measure has been applied to the two groups.

- Goodall1 similarity of the set of named entities found in the title and summary
- Goodall1 similarity of the set of named entities found in the article's body
- Jaccard coefficient of the set of named entities found in the title and summary
- Jaccard coefficient of the set of named entities found in the article's body

One-sentence similarity features

Due to the importance of the headlines in the news articles (Conboy, 2007), they had to be considered when considering more sophisticated representations. In this case, the word embeddings models BETO and FastText were used for converting the headlines into vectorial representations, and Cosine similarity and Word mover's distance have been applied respectively.

- Cosine similarity of BETO embedding of the headline
- Word mover's distance of FastText embedding of the headline

Document similarity features

The third feature considered for each pair of documents is a popular one, cosine similarity of TF-IDF, which has been widely used for ranking (Yogish et al., 2020) as well as classification of documents of different lengths (Huang, 2008). Moreover, it is the only measure that considers all the text content of both articles at once.

- Cosine similarity of TF-IDF considering all the parts of the article

Meta-data similarity features

The last two features consider the properties of the news articles rather than their content. Therefore, they add new information to the feature matrix rather than representing articles' content in different ways.

- Absolute difference of days between the publication of both articles
- Absolute difference of number of terms in both articles

3.3.4 Pairwise probability

The classification model used for this task is a supervised logistic regression, trained with about 35% of all possible pairs of combinations of articles labeled manually.

The original goal of this section was to binary classify whether each pair of articles are about the same sexual violence case. However, by doing so, we would lose the similarity perception described by the features and we would end up with a binary output (either a 0 or a 1). Having a dichotomous output forced us to naively cluster the articles, which resulted in a bad clustering (even if we change the cut-off[4] of the classifier).

To preserve the similarity perception and not restrict the study to a binary output, the logistic regression model was used to predict the probability of belonging to class 1 (both articles are about the same case).

3.3.5 Clustering

Clustering is defined as the task of grouping similar objects. In the context of the present work, we aim to cluster news articles that present information about the same case.

The clustering algorithm chosen is Agglomerative Hierarchical Clustering. It groups objects to form a binary tree starting from singleton[5] clusters; at each iteration, the most similar clusters are merged depending on their distance score (Ghoshdastidar et al., 2018).

---

[4] Threshold value used by a classifier to determine whether it belongs to class 0 or 1. Typically, the default threshold of most models is 0.5.

[5] Clusters that contain only one element.

There are three main linkage modalities determining the distance considered for merging two clusters (Murtagh & Contreras, 2017):

- Single linkage: considers the minimum distance between all elements of the two clusters.
- Complete linkage: considers the maximum distance between all elements of the two clusters.
- Average linkage: considers the average distance between all elements of the two clusters.

The linkage modality selected was 'average', and the distance matrix that the algorithm works upon was given by the N x N symmetric matrix where each position is computed with the expression below.

$$distance_{i,j} = 1 - prob(case_j = case_i) \tag{7}$$

N being the total number of articles, i and j being two articles. The probability that i and j are about the same case is given by the prediction estimated with a logistic regression model as detailed in the previous section—*3.3.4 Pairwise probability*.

The performance of the clustering can be assessed by with the Fowlkes-Mallow index, which evaluates the similarity between two clustering—the ground truth and the predicted one—as a combination of True Positives (TP), False Positives (FP), and False Negatives (FN); it does not consider True Negatives (TN) because TN denotes the number of points in the dataset that do not belong to the same cluster and that were neither grouped by the algorithm (Campello, 2007).

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \tag{8}$$

The output range is $[0, 1]$, 0 being the worst clustering possible and 1 being the best one.

3.4 Articles analysis

The present work has two different goals of analysis of the news articles about sexual violence. Firstly, to study the coverage of cases of sexual assault in the media considering three different characteristics: the type of sexual violence crime, the relationship between the victim and the

perpetrator and the place where the crime took place. Secondly, to get a broader view of the type of information typically presented in articles covering sexual assault crimes, we check whether distinct types of information are mentioned.

The approach followed in both cases is based on the assumption that by checking the presence and absence of terms and regular expressions, we can deduce whether some information is present in an article. The terms and regular expressions used for the analysis are presented in *ANNEX 2. Terms and regular expressions used for articles analysis*.

3.4.1 Coverage of cases analysis

The comparison between the coverage of sexual violence cases by media outlets and the official information released by Instituto Nacional de Estadística and Ministerio de Igualdad was done by comparing three different characteristics: the type of sexual violence (sexual assault, sexual harassment, or sexual abuse), the bond between the victim and the person who committed the crime (relationship bond, familiar bond, they knew each other, or they were strangers), and the place where the crime occurred (in a house, a public place, workplace, educational place or leisure-type of place).

Figure 6: Pipeline followed for coverage of cases analysis

The first step was to define different terms and regular expressions which can be associated with the aspects aimed to analyze. Subsequently, we checked whether some of these regular expressions appeared in each article. E.g., to know if an article mentions a familiar bond, terms and regular expressions established for the category "relative bond" (such as "daughter", "son", and "grandparent") were queried.

Afterward, all the articles presenting information about the same case were grouped—considering the clusters described in section *3.3 Cases' classification*—and the characteristics of each cluster were inferred by taking into account the presence and absence of regular expressions in every article of the cluster. However, due to the distinct nature of the aspects studied—type of violence, the victim-perpetrator relationship, and the place where the crime happened—their occurrences in the articles cannot be studied in the same way; therefore, different rules were defined for determining the most likely categories of each case.

The final step of the pipeline is the comparison between the results obtained and the official statistics provided by Instituto Nacional de Estadística (2020) and Ministerio de Igualdad (2020).

Sexual violence

Sexual violence cases can be classified into three types of crimes: sexual harassment, sexual assault, and sexual abuse—for further description, refer to section *2.1 Sexual violence statistics in Spain*.

For each type of sexual violence, we defined a list of terms and regular expressions presented in *ANNEX 2. Terms and regular expressions used for articles analysis*. For assigning a type of sexual violence to a case, we took into account the presence or absence—as a dichotomous variable—of the terms and regular expressions defined for each type of sexual violence in all the articles of the cluster. E.g., a cluster containing 5 articles has 3 binary inputs per article, each determining whether the article contains terms related to sexual assault, sexual harassment, and sexual abuse.

Having checked the presence of regular expressions in every article of the dataset, we assessed the type of sexual violence that most characterizes each case as detailed below:

- A case is classified as harassment if it is the only type of sexual violence that appears in at least one article of the cluster, i.e., neither sexual assault nor sexual abuse-related terms have been found in any of the articles of the cluster.

  This assumption is inferred from the definitions of the types of sexual violence since harassment is the only type that does not involve penetration; if there is any term or expression implying penetration, the case should be classified as sexual assault or abuse.

- If the articles of a cluster contain harassment and assault-related terms, the cluster of articles is classified as sexual assault.

- If the articles contain information related to sexual abuse, the cluster is classified as sexual abuse. Sexual abuse cases are defined as unwanted physical contact without violence—by definition—because the victim is not able to defend themself because they are in a situation of power imbalance; therefore, if a vulnerable victim is detected, then the type of sexual violence should be sexual abuse.

For the application of these rules, one variable per type of sexual violence is needed—indicating whether any of the articles of the cluster contains information related to the corresponding type. Given the three variables of all the articles of a cluster, the final variables are created using the Inclusive OR logic operation.

Example 1:Creation of per-case variables for determining the type of sexual violence

A cluster containing 4 articles will have as input 12 variables (3 per article):

- Article 1: {harassment = 0, assault = 1, abuse = 1}
- Article 2: {harassment = 0, assault = 0, abuse = 1}
- Article 3: {harassment = 0, assault = 0, abuse = 1}
- Article 4: {harassment = 0, assault = 1, abuse = 1}

These 12 variables will be converted into three by means of Inclusive OR operations:

- Harassment = 0 or 0 or 0 or 0 = 0
- Assault = 1 or 0 or 0 or 1 = 1
- Abuse = 1 or 1 or 1 or 1 = 1

The final variables considered for the classification of this case are:

- {harassment = 0, assault = 1, abuse = 1}

Having three dichotomous variables, the previous assumptions are used for determining the most suitable type of sexual violence for each cluster. The following table represents all their possible combinations and the corresponding results according to the rules described previously.

Table 7: Type of sexual violence, binary classification rules

INPUT			**OUTPUT**
Harassment	**Assault**	**Abuse**	**Classification**
0	0	0	Not classified
0	0	1	Abuse
0	1	0	Assault
0	1	1	Abuse
1	0	0	Harassment
1	0	1	Abuse
1	1	0	Assault
1	1	1	Abuse

Victim-perpetrator bond

The deduction of the type of bond between the victim and the perpetrator requires a distinct approach since news articles do not present information strictly about the crime, as they tend to include additional information about the case. Therefore, it is not straightforward to deduce the bond between the victim and the perpetrator as more than one type of bond may appear in articles.

The categories of relations considered are the following ones:

- Relationship: the victim and the criminal were in an affective-sexual relationship when the sexual violence occurred or had been in a relationship before.
- Relative: the victim and the perpetrator have a familial bond, they are relatives.
- Friend or acquaintance: whether the victim and the criminal knew each other before the crime, considering, among others, friends, teachers, workmates.
- Stranger: when the people involved in the sexual violence were strangers.

In this case, we defined three lists with terms and regular expressions, one for "relationship", another for "relative", and another for "acquaintance"; we did not create any list for the "stranger" category because it is difficult to detect using queries whether someone is a stranger, as it is a broad concept. The lists used can be found in *ANNEX 2. Terms and regular expressions used for articles analysis*.

To determine the relationship between the perpetrator and the victim, we considered the total number of articles mentioning each category rather than dichotomous variables. While articles may include information about people that are neither the victim nor the perpetrator, we can safely assume that the focus of these articles is on either the perpetrator or the victim, therefore, their bond should be the most frequent one.

The rules used for selecting the most appropriate bond are the following ones:

- If there is not any occurrence with any regular expression, define the relationship as "stranger".
- If there is only one category with a counter larger than 0, assign that category to the cluster.
- If there are multiple categories with counters larger than 0, but there is only one maximum, assign the maximum category to the cluster.

  For the creation of this rule, we are assuming that (1) not all the articles present the same information and (2) the number of occurrences of the bond between the victim and the criminal predominates any other secondary information presented by the media.
- If there is not one unique maximum—e.g., the 'relative and 'acquaintance' categories are mentioned by all the articles of one cluster—compute the number of occurrences considering only the headlines of the articles of all the categories and check again with the same rules.

Place of the crime

The last attribute inferred for each cluster of articles is the place where the sexual violence crime took place. The different types of places considered are:

- Public: multiple types of public places were considered, such as streets, avenues, public transport stops or inside a public transport vehicle, beaches, mountains, parks.
- Workplace: the place where someone works, such as an office, a shop, a coworking space, a factory.
- House: a place where people live, residences.
- Educational: a place where educational activities occur, such as schools, libraries, high schools, music schools, universities.
- Leisure: recreation places such as cinemas, theaters, bars, restaurants, shopping centers, nightclubs.

ANNEX 2. Terms and regular expressions used for articles analysis presents the regular expressions used for determining whether articles include information about the aforementioned types of places.

Having computed the number of occurrences of the different types of places in the articles of each cluster, we assumed that the place where the sexual violence crime occurred should be the most predominant one. However, some types of places—like public places—may have a lot of presence in articles without necessarily being where the crime happened.

The process of selecting the most likely place was done in steps: stop when determining the type of place, otherwise, execute the next step.

- **Dichotomous rule**: Count the number of articles in which each type of place is mentioned and select the most mentioned one.

  If there are two or more types of places appearing in the same number of articles, go to the next stage.

- **Headline and subtitle frequency rule**: Select the type of place with the highest frequency in the lead of the articles—the headline and subtitle—considered. Likewise, if there is only one type predominant, select it; else, go to the next stage.
- **Body frequency rule:** Select the type of place with the highest frequency considering the body of the articles.

- Finally, if any of the three steps did not output a type of place, leave it as unknown.

The definition of rules assumes that the most relevant information is placed in the headline and subtitle of an article, while the body includes more details (Conboy, 2007).

The output of the classification will be compared with the results of the *Macroencuesta 2019* (Ministerio de Igualdad de España, 2020), the latter contains information about the places where crimes happened only for those cases in which the victim and the perpetrator did not have a relationship type of bond; therefore the place characteristic was checked only for those clusters where the bond is assessed to be "relative", "acquaintance" or "stranger".

3.4.2 Content analysis

How articles present information depends directly on the characteristic writing style of the author, making it a complex task to study. Consequently, the present analysis aims to obtain quantitative information about the presence or absence of some types of information in the dataset and their location in each article—headline, subtitle, or body of the article. Considering all the articles of the dataset, we can infer which are the types of information that usually appear in the articles and whether there is an association between them.

The aspects examined have been classified into three categories: general information about the case, general stigmas that the society has about victims and perpetrators, as well as personal expression.

In a similar way as before, different terms and regular expressions have been defined for each feature—based on section *2.2 Representation of sexual violence in the media*—and they have been queried to every article in the dataset. However, in this case, we are not interested in the frequency of each regular expression but rather in their presence and absence. *ANNEX 2. Terms and regular expressions used for articles analysis* contains detailed information about the terms and regular expressions that were used for both section *3.4.1 Coverage of cases analysis* and *3.4.2 Content analysis*.

It is crucial to notice that even if some features are the same as in the previous section, it is a distinct type of analysis since we look at articles as individual publications rather than clusters.

Case-related information

This category is used to detect which type of information is frequently presented in articles about sexual violence and whether there is an association between them. The features checked are the following ones:

- Sexual violence acts: whether they tend to describe the type of sexual violence and the actions done by the perpetrator.
- Age: if information about the age of the people involved is typically mentioned.
- Time: if the part of the day in which the crime happened is usually pointed out.
- Bond: whether it is a common thing to specify the relationship between the people involved.
- Place: if articles provide information about the places where the sexual violence crime occurred.

Stigmas and expression

Media and news articles have always been powerful tools for mass manipulation (Fitzpatrick, 2018). Therefore, stigmatized information in articles about sexual violence cases can influence the way people think about the victims, the perpetrators, and the situations in which these crimes occur. This section tends to detect whether some general stigmas are present in the news articles of the data collection:

- Origin: whether some types of information about the origin of a person is present in the article.
- Intoxicated: if the article contains terms related to an intoxicated state and the presence of terms related to alcohol and other types of drugs.
- Clothing: terms involving the clothing of someone.
- Vulnerability: whether articles contain terms that referring to someone in a vulnerable state, such as alone, young, minor, abandoned.
- Aggressor: stigmas about the perpetrator of sexual violence crimes, by the presence of terms attributing mental illness or characterizing someone as a sexual predator, as a narcissist, and so on.

The writing style of an author shows the intention and reliability on the data related. Therefore, we also studied whether expressions that show a lack of confidence or euphemism are widely used in sexual violence news articles

- Doubt: if one can usually find expressions showing incredulity such as ‘alleged perpetrator’, ‘presumed crime’ and ‘presumed assault’, among others.

- Euphemism: if the reporters typically use mild or vague expressions that soften the harshness of sexual violence.

Remember that we cannot know if the information is about the victim, the perpetrator, or other people involved.

3.4.3 Association rules

The results from the previous parts can be further analyzed by means of association rules. Association rules are one of the most relevant techniques in data mining that allow us to discover patterns, relevant item sets, and associations between items in our dataset (Wei et al., 2009). Association rules are composed of two parts: $X \rightarrow Y$, X being the antecedent set and Y the consequent set, stating that the presence of the antecedent implies the presence of the consequent.

There are three main metrics used for the assessment of association rules:

- Support: $\sup(X)$ represents the relative frequency of an item set[6] in the dataset.

$$sup(X) = \frac{freq(X)}{N} \tag{9}$$

$$sup(X \rightarrow Y) = \frac{freq(X,Y)}{N} \tag{10}$$

- Confidence: $conf(X \rightarrow Y)$ is the conditional probability that Y occurs given X

$$conf(X \rightarrow Y) = \frac{freq(X,Y)}{freq(X)} \tag{11}$$

- Lift: $lift(X \rightarrow Y)$ is the ratio between the observed support and the expected support if X and Y were independent.

$$lift(X \rightarrow Y) = \frac{sup(X,Y)}{sup(X)sup(Y)} \tag{12}$$

When searching for frequent patterns and association rules, these metrics are used for determining their relevance by setting minimum thresholds. The most frequently used

[6] Set containing one or more elements of the dataset.

algorithm for mining association rules is an improved version of Apriori, which scans the dataset once and identifies the most frequent item sets of different sizes (considering a support threshold) and the possible association rules composed with them; the output is a list of association rules with their corresponding support, confidence and lift, for further interpretation (Yuan, 2017).

For the acquiring of meaningful associations, it is crucial to filter out those associations given by the elements' definitions or negative correlations that happen when the consequent is more likely to happen than the antecedent and the consequent together—i.e., the rule $X \rightarrow Y$ should not be considered if $\sup(Y) > conf(X \rightarrow Y)$—(Kabir et al., 2015).

In the present paper, association rules were used to compute the relationship between the characteristics extracted in sections *3.4.1 Coverage of cases analysis* and *3.4.2 Content analysis*.

Adapting the terminology to our data, N is the total number of articles, and the items are the previously extracted characteristics such as "sexual assault", "sexual violence", "sexual harassment", "relative bond", etc.; therefore, the frequency of an itemset is the number of articles that contain it.

Given the association rules extracted with the improved version of Apriori, we can reveal insights about the information that is usually presented together in articles describing sexual violence offences. E.g., exploring if sexual violence types have any relationship with the bond between the victim and the perpetrator, or with the place of the crime.

After all, association rules provide intrinsic correlations and co-occurrences of the data that cannot be obtained without a mining algorithm.

4 RESULTS

4.1 Data collection

4.1.1 Twitter scraping

The first phase of the project consists in the creation of the dataset. Starting with the collection of 10,000 tweets from media outlets' Twitter accounts—refer to *Table 4* for further description.

The scraping process provides only general information about each tweet collected—the tweet id, the source link of the tweet, the username of the account that published the tweet, the text of the post, the publication date, the number of retweets and favorites, and the mentions and hashtags

Collecting a concrete number of posts for all media outlets implies having different time ranges due to their different posting habits.

The data collection phase directly impacts the data analyzed and, consequently, the results too. Therefore, it is essential to consider the time ranges that our dataset captures for each media outlet and to highlight that the only time period that coincides for all the accounts is from 2020-07-24 to 2020-09-16, as can be seen in *Table 8*.

Table 8: Time ranges of the dataset collected for each media outlet

MEDIA OUTLET	TWITTER ACCOUNT	START DATE	END DATE	DAYS
20 Minutos	20m	2020-07-24	2020-09-16	54
El País	el_pais	2020-06-10	2020-09-16	98
LaSexta	sextaNoticias	2020-05-08	2020-09-16	131
El Confidencial	elconfidencial	2020-06-17	2020-09-16	91
RTVE News	rtve	2020-05-17	2020-09-16	122
LaVanguardia	LaVanguardia	2020-06-30	2020-09-16	78
ABC	abc_es	2020-07-21	2020-09-16	57
Cadena SER	La_SER	2020-05-29	2020-09-16	110
OK Diario	okdiario	2020-07-07	2020-09-16	71
El Diario	eldiario	2020-08-28	2020-09-16	19
El Periódico	elperiodico	2020-07-02	2020-09-16	76
La Razón	larazon_es	2020-07-23	2020-09-16	55
Telecinco	informativost5	2020-04-30	2020-09-16	139
El Mundo	elmundoes	2020-06-23	2020-09-16	85
Antena 3 online	A3Noticias	2020-05-29	2020-09-16	110

Since the data was obtained by retroceding in the timeline, the end date is the same for all the media outlets, and the start date is the publication date of the last tweet collected, thus the oldest.

In *Figure* 8, we can see with greater detail the posting habits of each Twitter account represented with boxplot diagrams the number of days it takes each account to post 1,000 tweets; providing us with a better perception of the regularity of the posting habits of the media outlets considered, and how the tweets scraped are distributed in their corresponding timeline.

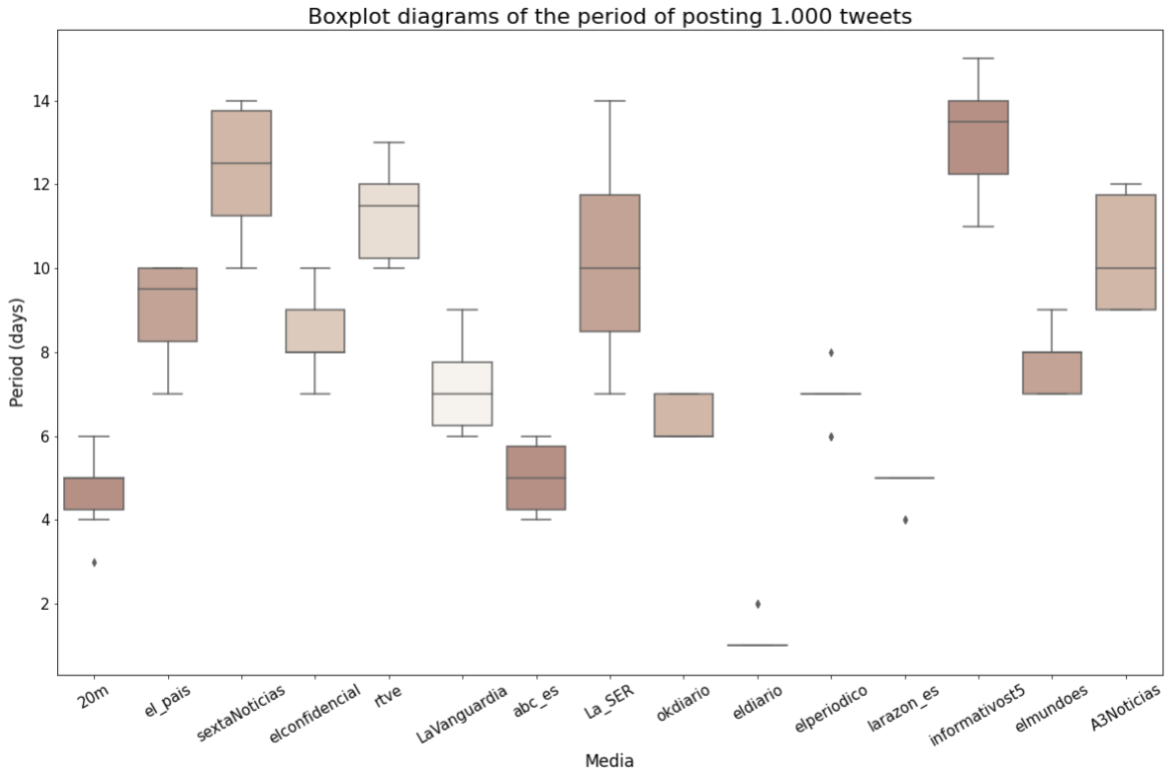

Figure 7: Boxplot diagrams of the period of posting 1,000 tweets for each media outlet

Patterns related to the posting frequency of the outlets can be observed: the average posting frequency is estimated to be a week and a half for 1,000 tweets, with the exception of two groups of outliers: the first one being characterized by a uniform high posting frequency—like El Diario, 20 Minutos and La Razón— and the second one by a wide posting time range—as El Confidencial.

4.1.2 Twitter hydration

The hydration process is prone to reduce the size of the data since some tweets may have no longer be available, as their authors erased them in between the scraping and the hydration processes.

Table 9: Tweets lost during hydration per Twitter account

TWITTER ACCOUNT	TWEETS LOST DURING HYDRATION	TOTAL NUMBER OF TWEETS
20m	3	9,997
el_pais	5	9,995
sextaNoticias	2	9,998
elconfidencial	4	9,996
rtve	1	9,999
LaVanguardia	1	9,999
abc_es	1	9,999
La_SER	2	9,998
okdiario	0	10,000
eldiario	3	9,997
elperiodico	2	9,998
larazon_es	3	9,997
informativost5	2	9,998
elmundoes	1	9,999
A3Noticias	2	9,998
TOTAL	32	149,968

As shown in *Table 9*, only 32/150,000 tweets were lost during the hydration of the tweets scraped, and none of the media outlets was significantly affected. The average of tweets lost is 2.13 per media, representing a 0.02% of the total collection.

4.1.3 Tweets classification

Tweets were classified by means of a supervised logistic regression model to distinguish those deemed to describe sexual violence cases.

As explained beforehand—in the *Data labeling* section of point *3.2.3 Tweets classification*—the labeling of the tweets was done in an iterative fashion using the classifier for speeding up the process. The tables below—*Table 10*, *Table 11*, and *Table 12*—show the evolution of the performance of the classifier as the number of manually labeled tweets increases.

The first version of the classifier was trained with 30,000 tweets, which were manually labeled, i.e., 2,000 tweets per media outlet. The main challenge of the training was the low portion of positively labeled tweets; among the first 30,000 tweets, only 192 are related to sexual violence.The train-test split was 70%-30% respectively.

Table 10: Performance of the initial model for the classification of sexual violence themed tweets

METRICS	LABEL 0 (NO SEXUAL VIOLENCE)	LABLE 1 (SEXUAL VIOLENCE)
Precision	0.99	1.00
Recall	1.00	0.20
F1-score	0.99	0.33
Accuracy	0.99	
Support	8830	35

Each classification metric assesses the performance of the model in a different way. Given a label X, its precision captures the proportion of records correctly classified as X among the total number of records classified as X, i.e., $TP/(TP + FP)$; in contrast, recall captures the portion of tweets correctly classified as X among all those whose ground truth is X, i.e., $TP/(TP + FN)$. Focusing on the metrics for label 1 in *Table 10*; precision = 1 means that all the tweets classified as 1 had indeed the label 1, while recall = 0.2 means that 80% of the tweets whose ground truth is 1 were wrongly classified by the model. Therefore, we are interested in maximizing the recall of the classifier.

Having labeled 5,000 tweets per media outlet (a total of 75,000 tweets), the portion of posts positively labeled was 0.55%—i.e., 409 tweets. However, the recall increased to 0.4, as can be seen *Table 11*, below.

Table 11: Performance of the classification of sexual violence themed tweets, training data size considered: 75,000 tweets

METRICS	LABEL 0 (no sexual violence)	LABLE 1 (sexual violence)
Precision	0.99	0.86
Recall	0.99	0.40
F1-score	0.99	0.55
Accuracy	0.99	
Support	17538	30

Following this approach, we labeled 70% of the dataset—105,000 tweets out of 150,000—identifying a total of 587 tweets deemed to describe sexual violence. With this data labeled, we trained the last version of the classifier, which provided a recall equal to 0.8—i.e., from all the tweets about sexual violence, only 20% of them were wrongly classified.

Table 12: Performance of the classification model for detecting sexual violence-themed tweets

METRICS	LABEL 0 (no sexual violence)	LABLE 1 (sexual violence)
Precision	1.00	0.86
Recall	1.00	0.80
F1-score	1.00	0.83
Accuracy	1.00	
Support	30514	400

The classifier labeled positively 867 tweets. Considering model's performance, it was probable that some of the tweets about sexual violence were not correctly classified; therefore, we manually checked the labeling of the 1,000 tweets with a high probability of being about sexual violence. The final number of tweets referring to sexual violence crimes is 883. The rest of the tweets—149,117—were no longer used.

4.1.4 News articles scraping

The last step for the creation of the dataset is the collection of articles from media outlets' webpages, starting from collecting the URLs attached to each tweet detected to be about sexual violence.

Since the classification was done with the summary of the tweets, which do not contain the URLs attached, each tweet was matched with their extended version, and we obtained 458 links.

However, 105 tweets included the URL with a 'nofollow' tag. Since these links do not appear in the extended version of the tweet, they were obtained by means of a chromium browser—for further details, check the *Limitations* section from point *3.2.4*.

The remaining tweets—304—ended up not being used as they either did not contain a URL, or the web address attached pointed to a webpage of no interest for the present study. Unfortunately, this was the case of all the tweets published by media outlets El Pais—@el_pais— and El Diario—@eldiario—which accumulated a total of 159 tweets about sexual violence that had to be discarded.

The total number of URLs collected was 563, without considering duplicates coming from retweets or different tweets containing the same web addresses attached. By following the unique set of URLs and scraping media outlets' websites, we obtained the final articles collection: 496 articles about sexual violence in NewsML-G2 format.

Table 13: Summary of Data collection process by media outlet

MEDIA	SEXUAL VIOLENCE TWEETS	URLS COLLECTED	DUPLICATED ARTICLES	FINAL UNIQUE ARTICLES
20m	76	75	11	64
el_pais	96	0	0	0
sextaNoticias	78	45	0	45
elconfidencial	39	21	3	18
rtve	23	8	0	8
LaVanguardia	15	13	0	13
abc_es	67	67	4	63

La_SER	46	36	5	31
okdiario	31	12	1	11
eldiario	63	0	0	0
elperiodico	49	36	8	28
larazon_es	45	27	1	26
informativost5	103	97	3	94
elmundoes	87	69	20	49
A3Noticias	65	57	11	46
TOTAL	883	563	67	496

The distribution of the articles obtained among the media outlets is not uniform; 54.43% of the articles were published by only four sources: Informativos T5, 20 Minutos, ABC Diario and El Mundo. The rest of the articles belong to the eight remining outlets.

4.2 Cases' classification

The classification of news articles into cases was performed in two steps: the prediction of the similarity probability for each pair of articles and the clustering of articles given the output of the previous step.

4.2.1 Pairwise features

To predict the probability of a pair of articles being about the same case, we used the features detailed in section *3.3.3 Features*. *Table 14* shows the correlation between the features considered, helping us understand the way in which they are related. As it can be seen, there are only two pairs of features with a correlation higher than 0.5—in absolute terms—meaning that most of them are able to represent articles' information in distinct ways, and it makes sense to include them as independent features.

Table 14: Correlation of the features used for computing the similarity between pairs of articles

	Goodall1 articles' leads	**Goodall1 articles' bodies**	**Jaccard articles' lead**	**Jaccard articles' bodies**	**TF-IDF full article**	**BETO headlines**	**WMD headlines**	**Difference dates**	**Difference n. terms**
Goodall1 articles' leads	**1.00**	0.07	**0.85**	0.34	0.40	0.10	-0.07	0.00	-0.02
Goodall1 articles' bodies	0.07	**1.00**	0.04	0.20	0.12	0.03	-0.02	0.00	0.01
Jaccard articles' leads	**0.85**	0.04	**1.00**	0.34	0.37	0.09	-0.06	0.00	-0.02
Jaccard articles' bodies	0.34	0.20	0.34	**1.00**	**0.57**	0.15	-0.08	-0.01	-0.09
TF-IDF	0.40	0.12	0.37	**0.57**	**1.00**	0.25	-0.15	-0.01	-0.03

BETO headlines	0.10	0.03	0.09	0.15	0.25	**1.00**	-0.24	0.04	-0.15
WMD headlines	-0.07	-0.02	-0.06	-0.08	-0.15	-0.24	**1.00**	-0.03	0.02
Difference dates	-0.00	0.00	-0.00	-0.01	-0.01	0.04	-0.03	**1.00**	-0.02
Difference n. terms	-0.02	0.01	-0.02	-0.09	-0.03	-0.15	0.02	-0.02	**1.00**

The main relations between the features, highlighted in the table above, are:

(1) The Jaccard coefficient of the lead and the Goodall1 similarity of the lead are highly correlated: 0.85. This can be explained from their computation: both measures are set-based and take as input the named entities extracted from the articles' lead. However, the same metrics computed with the named entities extracted from articles' bodies have considerably lower correlation, equal to 0.20.

Since metrics depend only on the entities extracted, the reason why both correlations are that different comes down to the article's content. On the one hand, articles' lead tends to provide precise information, being more likely to contain few and specific named entities; consequently, most of the pairs of articles will hardly share any entity—which is the case of 98.81% of the pairs—in this situation, both measures assign the same similarity value: 0.

On the other hand, the body of an article contains further information about the case, which may result in largest sets of entities. While Goodall1 estimates the similarity of two sets by considering their probability of co-occurring randomly, Jaccard coefficient only considers the portion of common entities; therefore, these metrics are more likely to present different results since they can, actually, be computed.

(2) Cosine similarity of TF-IDF and the Jaccard Coefficient of the named entities extracted from the body present a high correlation score, equal to 0.57. This relationship can be understood by contemplating the following information:

- The input data of these features overlaps for the major part: cosine similarity of TF-IDF considers all the article, and Jaccard Coefficient extracts the named entities from articles' bodies.
- Their output ranges are equal. Cosine similarity output range goes from -1 to 1; however, TF-IDF vectors do not contain negative entries and the final range is reduced to [0, 1], coinciding with Jaccard coefficient's range.

The rest of the features have presented low correlation scores in the range [-0.3, 0.3].

4.2.2 Pairwise classifier

The classifier used to distinguish the pairs of articles providing information about the same case was trained with 36,046 pairs of articles manually labeled—around 30% of the total number of possible pairs—and only 461 were about the same case.

Considering that the portion of positively labeled pairs was 1.27%, the performance of the logistic regression model—shown in

Table 15—is remarkable.

Table 15: Performance of the classification model for detecting pairs of articles about the same case

METRICS	LABEL 0 (different cases)	LABLE 1 (same case)
Precision	1.00	0.94
Recall	1.00	0.90
F1-score	1.00	0.92
Accuracy	1.00	
Support	7121	89

Figure 8 shows the importance of each feature deduced from the weights that the model assigns. The most relevant features observed are TF-IDF, and the Jaccard coefficient considering both articles' leads and bodies.

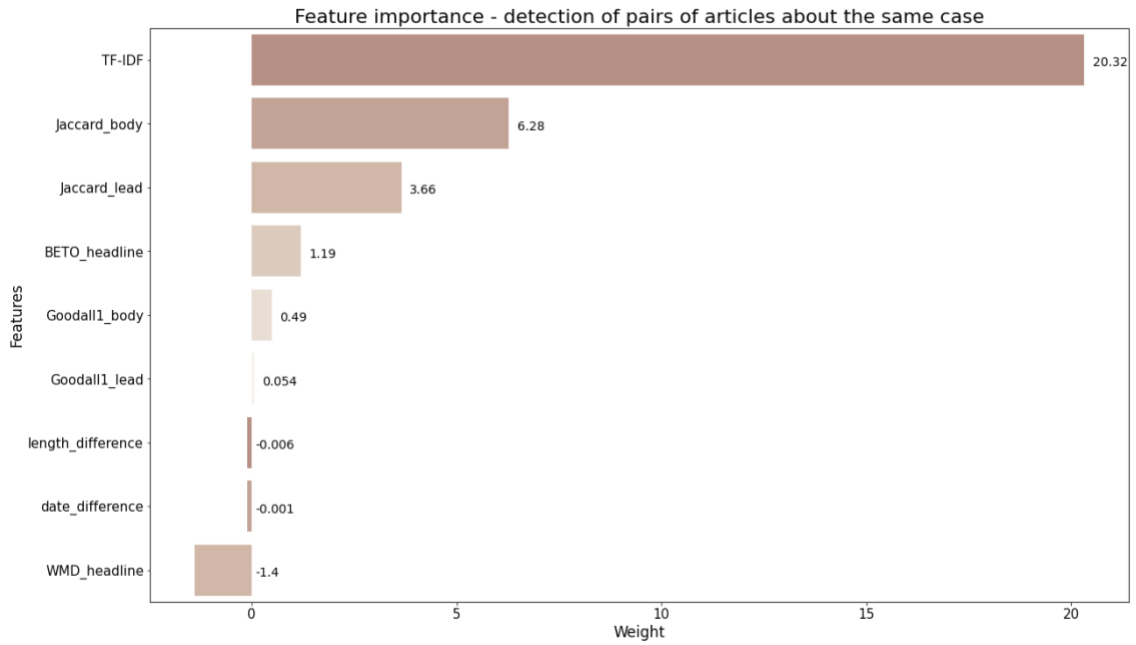

Figure 8: Feature importance of the pairwise classifier for detecting articles about the same case

Furthermore, the features that assess the distance between two documents, rather than the similarity—Word Mover's Distance, length difference, and difference of the publication dates—have been assigned negative weights by the model.

The trained classification model was used to predict the probability, for each pair of articles, of providing information about the same sexual violence case. Resulting in a symmetric matrix of dimensions NxN, N being the total number of articles considered (i.e., 496), where each position $[i,j]$ can be interpreted as the normalized similarity between articles i and j. Analogously, we computed the distance between articles i and j as $1 - similarity(i,j)$.

4.2.3 Clustering

Articles about the same case were grouped by means of an Agglomerative Hierarchical Clustering algorithm, which was fed with the precomputed normalized distance matrix. The parameters selected for the clustering were:

- Linkage type: 'average'
- Distance threshold: 0.3

Implying that two clusters are joined only if the average distance between all the points of both clusters is less or equal than 0.3.

The clustering classified the 496 sexual violence articles into 289 cases.

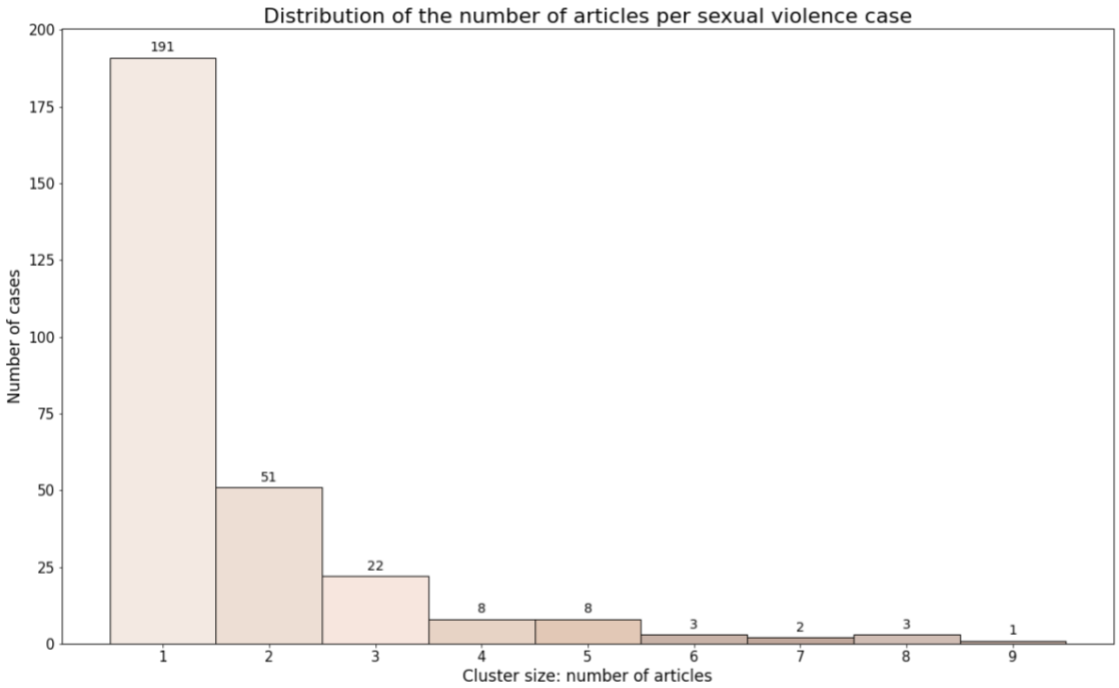

Figure 9: Distribution of the number of articles per sexual violence case – obtained by clustering articles

Figure 9 shows the distribution of cases in terms of the number of articles they contain. Only 8.65% of the cases had a high coverage—cluster size greater than 3—while 66.08% of the

cases have been explained in only one article. As a matter of fact, 25.2% of the articles (125) present information about 7.61% of the cases (22).

To evaluate the performance of the clustering, we computed the number of True Positives, False Positives, True Negatives, and False Negatives, considering as ground truth the pairs of articles labeled for the pairwise classification—refer to section *3.3.4 Pairwise probability* for more details.

Table 16: Evaluation of the clustering of articles

METRIC	DESCRIPTION		RESULT
	Ground truth	Clustering	
TP	Both articles about the same case	Both articles in the same cluster	337 = 0.93%
FP	Articles about different cases	Both articles in the same cluster	8 = 0.02%
TN	Articles about different cases	Articles in different clusters	35577 = 98.70%
FN	Both articles about the same case	Articles in different clusters	124 = 0.34%

The overall performance, computed with the Fowlkes-Mallows index—consult *(8)*— is 0.85. It is crucial to notice that most of the errors produced by the clustering algorithm are FNs, meaning that clusters hardly ever contain articles that do not share information about the same case as the rest of the articles in the cluster.

4.3 Articles analysis

4.3.1 Study group

Articles' analysis is based mainly on the presence and absence of terms and regular expressions associated with the attributes and characteristics studied. Therefore, it is essential to understand the quality of the search.

For this purpose, we selected a random sample of 50 articles—representing 10.08% of the entire collection—checked whether each of them mentioned the characteristics studied and compared the results with the outputs of the queries.

Table 17: TP, FP, TN, and FN definition for the query of terms and regular expressions

METRIC	DESCRIPTION	
	Ground truth	Query outputs
TP	Characteristic mentioned in the article	Characteristic detected by the query
FP	Characteristic not mentioned in the article	Characteristic not detected by the query
TN	Characteristic mentioned in the article	Characteristic detected by the query
FN	Characteristic not mentioned in the article	Characteristic not detected by the query

We computed TP, FP, TN, and FN for each element studied, following the guidelines shown in *Table 17*. Next, we used these metrics to calculate the precision, recall, and F1 score; the results of which can be found in *Table 18*.

Table 18: Performance of the query of terms and regular expression

ATTRIBUTES	MANUAL EVALUATION				METRICS		
	TP	FP	TN	FN	Precision	Recall	F1 score
Age	42	3	4	1	0.93	0.98	0.95
Sexual assault	31	3	16	0	0.91	1.00	0.95
Sexual harassment	18	2	29	1	0.90	0.95	0.92
Sexual abuse	29	2	18	1	0.94	0.97	0.95
Bond relationship	11	3	35	1	0.79	0.92	0.85
Bond relative	13	1	36	0	0.93	1.00	0.96
Bond acquaintance	29	2	18	1	0.94	0.97	0.95
Place public	32	9	9	0	0.78	1.00	0.88
Place workplace	10	1	39	0	0.91	1.00	0.95
Place house	19	2	28	1	0.90	0.95	0.93
Place educational	6	3	41	0	0.67	1.00	0.80
Place leisure	24	15	10	1	0.62	0.96	0.75
Time	8	1	38	3	0.89	0.73	0.80
Stigma intoxicated	9	0	39	2	1.00	0.82	0.90
Stigma clothing	6	1	41	2	0.86	0.75	0.80
Stigma origin	23	5	20	2	0.82	0.92	0.87
Stigma aggressor	3	0	46	1	1.00	0.75	0.86
Stigma vulnerability	32	2	26	0	0.94	1.00	0.97
Expression euphemism	9	0	36	5	1.00	0.64	0.78
Expression doubt	29	0	20	1	1.00	0.97	0.98

The F1 score is higher than 0.75 for all the categories studied, denoting an acceptable overall performance. Precision and recall help us understand the type of errors that each feature is most likely to contain in the results and to recognize the trustworthy range of each of them. A low precision denotes many FP values, whereas a low recall means a high number of FN.

4.3.2 Coverage cases

The coverage of cases has been studied by assigning to each case—clusters containing articles about the same crime—a type of sexual violence, a type of bond between the victim and the perpetrator, and a place, according to the presence of terms and regular expressions in the articles.

The following sub-sections detail the results found for each feature studied—i.e., the type of sexual violence is a feature, and the options are sexual assault, sexual harassment, and sexual abuse. It is important to notice that individual cases were classified but their respective articles were not; meaning that, the number of articles represents the quantity of news into the classified cases—i.e., if two cases containing 2 and 3 articles are classified as sexual harassment, then, the total number of sexual harassment articles is 5, without having explicitly classified the articles separately.

The results present for each feature studied and their corresponding options:

- A table with the number of cases and articles per option
- A table with the statistics of cluster's sizes per option
- A plot presenting the coverage of each option

Sexual violence

The coverage of different sexual violence types in our dataset was compared with the statistics provided by INE (Instituto Nacional de Estadística, 2020), which show the ratio of people convicted for each sexual violence type of crime during 2019.

Table 19: Number of cases and their corresponding articles per type of sexual violence

SEXUAL VIOLENCE	% INE 2019	Number of cases	Percentage of cases	Number of articles	Percentage of articles
Abuse	57.67%	181	62.63%	342	68.95%
Harassment	21.98%	24	8.30%	38	7.66%
Assault	20.35%	71	24.57%	101	20.36%
Not classified		13	4.50%	15	3.02%

Table 20: Summary statistics of cluster's sizes per type of sexual violence

SEXUAL VIOLENCE	Min size	Mean	Max size	Quantiles		
				0.25	0.50	0.75
Abuse	1,0	1.89	9,0	1,0	1,0	2,0
Harassment	1,0	1.58	6,0	1,0	1,0	2,0
Assault	1,0	1.42	5,0	1,0	1,0	2,0

The coverage of a case can be assessed by the number of articles broadcasting it. Consequently, we estimated the coverage of each type of sexual violence depending on the size of each cluster.

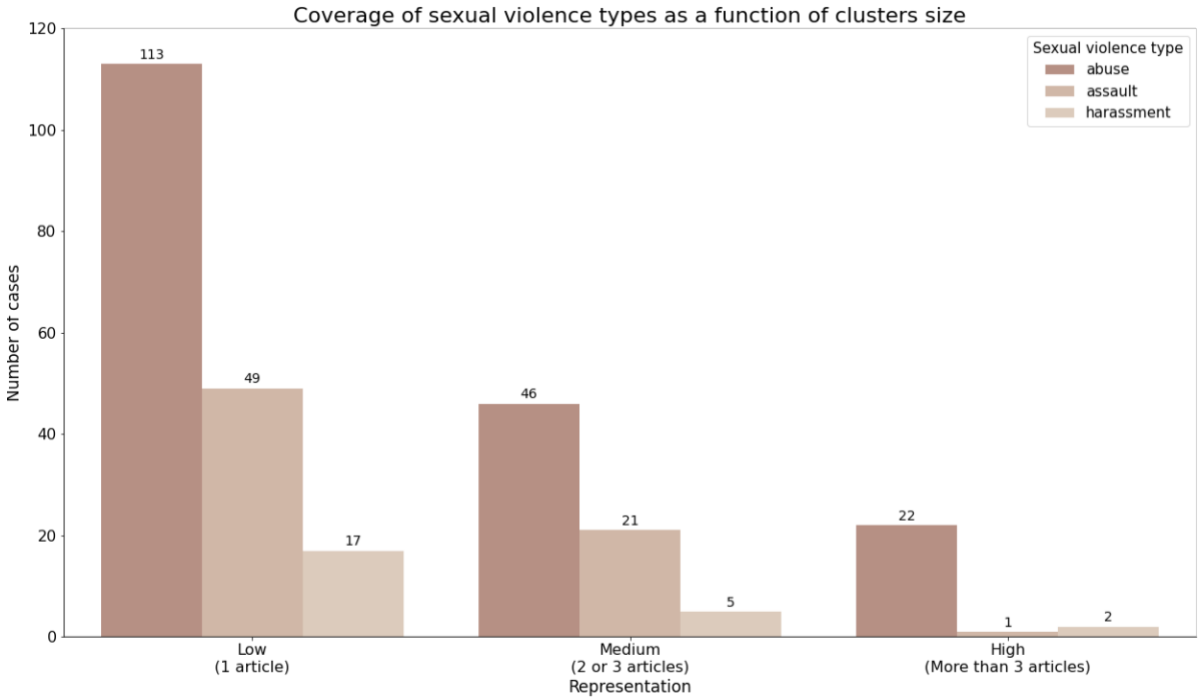

Figure 10: Representation of each sexual violence type as a function of cluters' sizes

Table 19 shows that harassment cases are less prevalent in the dataset than in the statistics considered, while the coverage of abuse and assault cases suggest that they are fairly represented. Besides, it can be appreciated from in *Figure 10* that the proportion of highly covered harassment cases is larger for abuse and assault.

Summary statistics of each type of sexual violence exposes that coverage does not largely differ in terms of quantiles, and differences in means result from variations on size of the maximum cluster.

Victim-perpetrator bond

The coverage analysis of the type of bond between the victim and the perpetrator was compared with the information provided by *Macroencuesta 2019* (Ministerio de Igualdad de España, 2020).

Table 21: Number of cases and their corresponding articles per type of bond between the victim and the perpetrator

BOND VICTIM PERPETRATOR	Macro-encuesta 2019	Number of cases	Percentage of cases	Number of articles	Percentage of articles
Relationship	72.20%	33	11.42%	66	13.31%
Acquaintance	14.03%	82	28.37%	150	30.24%
Stranger	10.86%	124	42.91%	211	42.54%

Relative	6.00%	39	13.49%	51	10.28%
Not classified	0.05%	11	3.81%	18	3.63%

It can be observed that most sexual violence crimes occur when there is a relationship-type of bond between the victim and the perpetrator. However, media outlets portray a different reality.

Table 22: Summary statistics of cluster's sizes per type of bond between the victim and the perpetrator

BOND	Min size	Mean	Max size	Quantiles		
				0.25	0.50	0.75
Relationship	1,0	2.00	7,0	1,0	1,0	2,0
Acquaintance	1,0	1.83	8,0	1,0	1,0	2,0
Stranger	1,0	1.70	9,0	1,0	1,0	2,0
Relative	1,0	1.30	3,0	1,0	1,0	2,0

Summary statistics provide further dimensionality to this data; we can observe that the "relationship"'s mean value is the highest one, implying that this type of bond has the largest proportion of highly covered cases (even if the maximum size of this category is not the largest one)

On the other hand, statistics also evidence that none of the cases in which the bond is "relative" has been highly covered by the media.

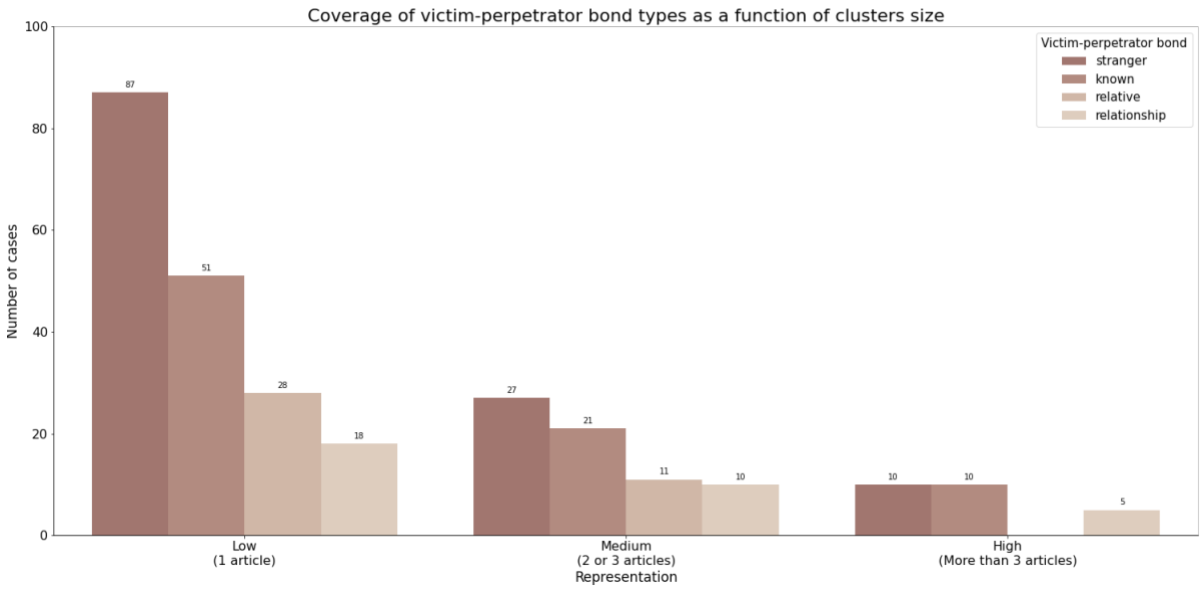

Figure 11: Representation of type of bond between the victim and the perpetrator as a function of cluters' sizes

Place of the crime

The third aspect studied in the coverage analysis is the place where the sexual violence crime took place. This information was compared with the data provided by *Macroencuesta 2019*

(Ministerio de Igualdad de España, 2020). However, *Macronecuesta 2019* only contains geographical information for those cases where the perpetrator and the victim did not have a relationship bond. Therefore, for the creation of the tables and figures below—*Table 23*, *Table 24*, and *Figure 12*—11.42% of the cases were not considered.

Table 23: Number of cases and their corresponding articles per place where sexual violence crimes happen

PLACE	Macro-encuesta 2019	Number of cases	Percentage of cases	Number of articles	Percentage of articles
House	65.4%	25	8.65%	36	7.26%
Public	36.3%	113	39.10%	196	39.52%
Leisure	15.6%	86	29.76%	145	29.23%
Workplace	7.1%	5	1.73%	6	1.21%
Education	6.2%	3	1.04%	14	2.82%
Not classified	0.3%	24	8.30%	33	6.65%

Table 24: Summary statistics of cluster's sizes per type of bond between the victim and the perpetrator

PLACE	Min size	Mean	Max size	Quantiles		
				0.25	0.50	0.75
House	1,0	1.44	3,0	1,0	1,0	2,0
Public	1,0	1.73	8,0	1,0	1,0	2,0
Leisure	1,0	1.69	9,0	1,0	1,0	2,0
Workplace	1,0	1.20	2,0	1,0	1,0	1,0
Education	2,0	4.67	8,0	3,0	4,0	6,0

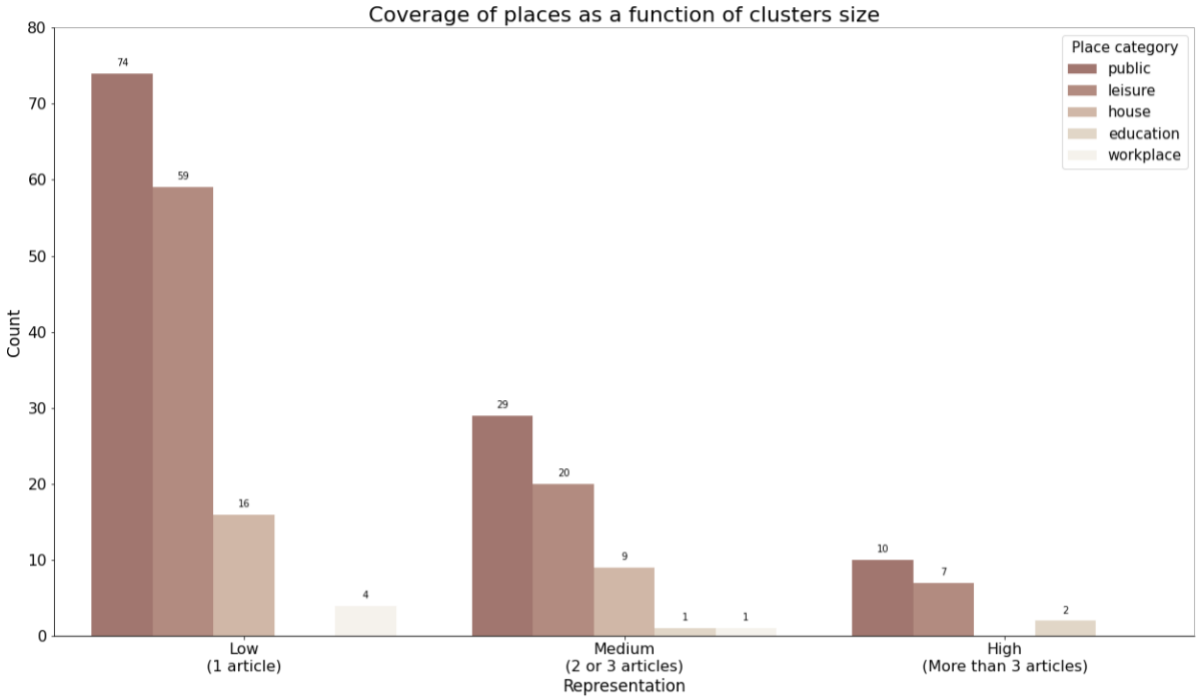

Figure 12: Representation of the places where sexual violence happens as a function of cluters' sizes

The main observation drawn from these results is that sexual violence occurs mostly in homes, but it is under-represented by the media, which deviates the attention from these places, accentuating and providing more coverage to sexual violent incidents that occur in leisure spaces. Additionally, the proportion of cases in our dataset happening in workplaces or educational places does not reach 2%.

4.3.3 Content analysis

To analyze which type of information is typically mentioned in sexual violence-related articles, we checked the presence of distinct attributes in the articles of the dataset. The analysis is focused on individual articles, and in which parts the examined aspects appear—headline, subtitle, or body of the article, presented in *Table 25.*

It is important to notice that even if most of the features were also evaluated in the previous section, the number of articles containing them are different, given that in the previous sections we were studying the cases.

Table 25: Content analysis results – presence of features in articles' headline, subtitle and body

TYPE	FEATURES	Total		Headline		Subtitle		Body	
		Articles	%	Articles	%	Articles	%	Articles	%
Sexual violence acts	Assault	358	72.18%	242	48.79%	151	30.44%	351	70.77%
	Harassment	201	40.52%	42	8.47%	39	7.86%	195	39.31%
	Abuse	310	62.50%	180	36.29%	96	19.35%	299	60.28%
Case-related information	Time	140	28.23%	2	0.40%	10	2.02%	139	28.02%
	Age	471	94.96%	229	46.17%	238	47.98%	470	94.76%
	Workplace	88	17.74%	5	1.01%	5	1.01%	87	17.54%
	Place: house	254	51.21%	10	2.02%	35	7.06%	252	50.81%
	Educational place	67	13.51%	13	2.62%	8	1.61%	65	13.10%
	Leisure place	386	77.82%	84	16.94%	74	14.92%	377	76.01%
	Public place	385	77.62%	152	30.65%	112	22.58%	380	76.61%
	Bond relationship	139	28.02%	38	7.66%	28	5.65%	139	28.02%
	Bond relative	162	32.66%	44	8.87%	47	9.48%	162	32.66%
	Bond acquaintance	291	58.67%	28	5.65%	38	7.66%	290	58.47%
Stigmas	Origin	283	57.06%	49	9.88%	56	11.29%	283	57.06%
	Intoxicated	92	18.55%	19	3.83%	11	2.22%	90	18.15%

	Clothing	62	12.50%	4	0.81%	6	1.21%	62	12.50%
	Perpetrator	20	4.03%	5	1.01%	0	0.00%	18	3.63%
	Vulnerability	378	76.21%	160	32.26%	142	28.63%	375	75.60%
Expression	Euphemisms	77	15.52%	1	0.20%	3	0.60%	76	15.32%
	Doubt	317	63.91%	70	14.11%	67	13.51%	306	61.69%

The characteristics studied are categorized into four groups—sexual violence acts, case-related information, stigmas, and expression—and the regular expressions used for examining whether they appear in articles or not are detailed in *2. Terms and regular expressions used for the analysis*.

While terms denoting harassment are present in 40.52% of articles, they are usually not the main focus of an article (only 8.3% of the articles focus mainly on sexual harassment), comprising the least prevalent type of sexual crime present in headlines.

There is a high attention directed towards the age of the people involved in the articles, since mentions of age can be found in 94.96% of articles, making it the most common information used in headlines. Moreover, the rest of the information studied tends to appear in articles' bodies, by being shown as extra information about the case rather than emphasizing it.

Stigmas continue being present in news articles without gaining importance in the headlines; the most prevalent ones are 'origin', 'vulnerability' and 'intoxicated', which tend to be attributed to the victim.

At the same time, doubt expressions prevail over euphemisms as they appear in 317 news articles of the collection, out of 496.

4.3.4 Association rules

The results obtained from the content analysis were further analyzed by means of association rules (AR). Using this data mining technique, we extracted two different groups of association rules considering (A) content-related features and (B) stigmas and expression. The thresholds used in both cases are minimum support = 0.1; minimum confidence = 0.5; and minimum lift = 1.05.

Due to the number of attributes considered in group A, we obtained 130 AR satisfying the aforementioned criteria. The top-15 association rules, selected considering the highest lift, are shown in *Table 26*. The full list of results can be found in *ANNEX 3. Association rules*.

Table 26: Top-15 case-related association rules sorted by lift

Antecedent		Consequent	Support	Confidence	Lift
Sexual abuse	Bond relationship	Bond relative	0.11	0.62	1.89
Place house	Bond relative	Bond relationship	0.11	0.50	1.80
Place public	Bond relationship	Bond relative	0.12	0.54	1.66
Sexual assault	Bond relationship	Bond relative	0.10	0.51	1.56
Place house	Sexual harassment	Bond relative	0.11	0.50	1.53
Sexual assault	Bond relationship	Place house	0.15	0.76	1.48
Place leisure	Bond relationship	Place house	0.18	0.75	1.47
Sexual abuse	Bond relationship	Place house	0.13	0.74	1.45
Age	Bond relationship	Place house	0.19	0.72	1.40
Place public	Bond relationship	Place house	0.16	0.71	1.38
Bond relationship		Place house	0.19	0.70	1.37
Place public	Bond relative	Place house	0.19	0.70	1.37
Place leisure	Bond relative	Place house	0.18	0.69	1.35
Bond relative	Sexual abuse	Place house	0.16	0.68	1.32
Sexual assault	Bond relative	Place house	0.15	0.67	1.31

By mining data the presence of terms denoting the presence of stigmas and expressions—group B—we obtained a total of 5 association rules; these can be seen in *Table 27.*

Table 27: Stigma and expression-related association rules

Antecedent		Consequent	Support	Confidence	Lift
Stigma intoxicated		Stigma origin	0.13	0.73	1.26
Stigma vulnerability	Stigma intoxicated	Stigma origin	0.11	0.70	1.23
Expression euphemism		Stigma vulnerability	0.14	0.88	1.16
Stigma intoxicated		Stigma vulnerability	0.16	0.84	1.20
Stigma vulnerability	Stigma intoxicated	Expression doubt	0.10	0.68	1.06

Finally, to study the relationship between all the informative features and the communicative ones, we computed the association rules containing both types of characteristics with more strict thresholds, in order to reduce the number of results and obtain more significant ones. Thresholds used were minimum support: 0.3, minimum confidence: 0.5 and minimum lift: 1.05. We obtained 40 association rules satisfying these conditions. *Table 28* contains the top-15 association rules sorted by lift.

Table 28: Top-15 inter-categories association rules sorted by lift

Antecedent		Consequent	Support	Confidence	Lift
Stigma origin	Sexual abuse	Place public	0.34	0.94	1.22
Expression doubt	Sexual abuse	Place public	0.37	0.91	1.17
Stigma vulnerability	Sexual abuse	Place public	0.48	0.90	1.16
Place public	Expression doubt	Bond acquaintance	0.34	0.67	1.14
Place leisure	Stigma vulnerability	Sexual abuse	0.43	0.71	1.14
Stigma vulnerability	Sexual assault	Sexual abuse	0.40	0.71	1.13
Place leisure	Expression doubt	Bond acquaintance	0.33	0.66	1.12
Place public	Stigma origin	Bond acquaintance	0.30	0.66	1.12
Stigma vulnerability		Sexual abuse	0.53	0.53	0.70
Stigma vulnerability	Age	Sexual abuse	0.53	0.69	1.11
Stigma vulnerability	Bond acquaintance	Expression doubt	0.31	0.70	1.10
Stigma vulnerability	Bond acquaintance	Place public	0.37	0.85	1.09
Expression doubt	Age	Bond acquaintance	0.39	0.64	1.08
Expression doubt		Bond acquaintance	0.41	0.41	0.63
Stigma origin	Stigma vulnerability	Place public	0.35	0.84	1.08

Complete results can be found in *3. Association rules*.

5 DISCUSSION

This section of the document discusses the findings which emerged from the development of the present project, focusing on the results presented in the previous chapter and answering the questions stated in section *1.3 Objectives*.

5.1 Dataset creation

The first defined goal was the creation of a dataset of news articles deemed to describe sexual violence cases, published by the most read online media outlets in Spain.

The collection of data was the first phase of the project. We started by selecting 10,000 tweets of the top-15 most read online media outlets in Spain. As shown in *Figure 7: Boxplot diagrams of the period of posting 1,000 tweets for each media outlet*, outlets do not follow any specific posting pattern; therefore, we have the same number of tweets for all outlets, but their time ranges differ depending on their posting frequency.

The collected tweets went under a hydration procedure, which is predisposed to lowering the number of tweets. However, our dataset was not compromised during the hydration since the quantity of posts lost is insignificant. The successfully hydrated tweets were binary classified, according to their relationship with sexual violence. The classification outputted 883 tweets concerned with the subject of interest, representing only 0.59% of the total number of tweets collected. Some classification examples can be found in *Example 2: Classification of tweets according to their relationship with sexual violence*. Positive and negative examples., below.

Example 2: Classification of tweets according to their relationship with sexual violence. Positive and negative examples.

Tweets positively labeled

He sexually abused his daughter from the age of nine to 16 when her mother went to work in Alicante. - Author: @abc_es. (Original: *Abusó sexualmente de su hija desde los nueve a los 16 años cuando la madre se iba a trabajar en Alicante)*

A 45-year-old man arrested in Valencia for showing his genitals to three minors. – Author: @informativost5. (Original: *Detenido un hombre de 45 años en Valencia por enseñar sus genitales a tres menores*)

After torturing and raping his victims, he abandoned them in the forest. – Author: @lavanguardia. (Original: *Tras torturar y violar a sus víctimas las abandonaba en el bosque*)

Tweets negatively labeled

A man was reported for hounding a group of dolphins with a jet ski. – Author: @informativost5. (Original: *Denunciado un hombre por acosar a un grupo de delfines con una moto acuática*)

How to be a pastor in the 21st century. – Author: @elmundoes. (Original: *Cómo ser un pastor en el siglo XXI*)

More than 40 million adolescents between the ages of 13 and 15 consume tabacco, acording to the WHO. – Author: @rtve. (Original: *Más de 40 millones de adolescentes entre 13 y 15 años consumen tabaco, según la OMS*).

The number of tweets about sexual violence per Twitter account had a high degree of fluctuation, and we did not observe any relationship between the quantity of positively labeled tweets and the posting frequency of each media outlet.

Nonetheless, from the results shown in *Table 13: Summary of Data collection process by media outlet*, we could detect a subtle correlation between the posting frequency and the proportion of unique articles about sexual violence: media outlets that post with a high frequency tend to post tweets linking to the same news articles—e.g., the outlets Informativos T5*,* Antena 3 and La SER. This behavior can be explained through the difficulty of bringing out many different articles in a short period of time (1 day).

Having labeled the tweets, the posts related to sexual violence were used as sources of news articles. During the process of collecting articles, we had to forego El Diario and El Pais, since we could not reach any URL pointing to online articles from their tweets.

According to Calvo et al. (2020), the final dataset, containing 496 articles, portrays a fair representation of the attention that outlets' pay to sexual violence articles, since their principal interest is to redirect users to their main websites.

5.2 Clustering of news articles

Once we had created the collection of articles, the second goal was to group articles that are concerned with the same case. We split the process into two steps: first, we assigned to each pair of articles a probability of being about the same case; secondly, we clustered the articles using the probabilities obtained in the first step as an assessment of their closeness.

The pairwise similarity score was assigned by means of a logistic regression model, which took as input diverse features computing the resemblance of each pair of articles. Features were created considering different text representation techniques and similarity metrics classified into four categories: set-based similarity, one-sentence similarity, document similarity and meta-data similarity—refer to section *3.3.3 Features* for a complete description.

Example 3 shows the two possible scenarios: the features comparing two articles about the same case, and the features comparing two articles about different cases.

Example 3: Pairwise features for articles both possible scenarios: articles about the same case vs articles about different cases.

Features for a pair of articles about the same case

Headline of article A[7]: A man has been arrested in Madrid accused of abusing his stepdaughter and using her for child pornography. (Original: *Detenido en Madrid acusado de abusar de su hijastra y utilizarla para pornografía infantil).*

Headline of article B[8]: Arrested for abusing his stepdaughter for child pornography for 7 years. (Original: *Detenido por abusar de su hijastra para pornografía infantil durante 7 años*).

Features

Set-based features				Document feature	One-sentence features		Metadata features	
Goodall1 lead	Goodall1 body	Jaccard lead	Jaccard body	TF-IDF	BETO headline	WMD headline	Dates diff.	Length diff.
0.00	0.01	0.00	0.38	0.75	0.85	0.66	0	1447

Features for a pair of articles about the different cases

Headline of article A[9]: One year of probation for the man who sexually assaulted a female reporter. (Original: *Un año de libertad condicional para el hombre que agredió sexualmente a una reportera*).

Headline of article B[10]: A gang, dedicated to the trafficking of women for sexual exploitation, dismantled in Mijas. (Original: *Cae en Mijas una banda dedicada a la trata de mujeres para su explotación sexual*).

Features

Set-based features				Document feature	One-sentence features		Metadata features	
Goodall1 lead	Goodall1 body	Jaccard lead	Jaccard body	TF-IDF full article	BETO headline	WMD headline	Dates diff.	Length diff.
0.00	0.00	0.00	0.00	0.01	0.68	1.01	292	289

A general property of these features is their small correlation—as presented in *Table 14: Correlation of the features used for computing the similarity between pairs of articles*—which denotes that each feature is able to represent the information in a unique way by emphasizing specific characteristics of the text. Nonetheless, two pairs of features present high correlations that can be explained from their computation similarities and the large number of True

[7] Article extracted from https://www.telecinco.es/informativos/sociedad/detenido-acusado-abusar-hijastra-pornografia-infantil_18_2970195132.html

[8] Article extracted from https://www.elmundo.es/espana/2020/06/28/5ef85c50fdddffa1648b45d2.html

[9] Article extracted from https://cadenaser.com/ser/2020/09/03/television/1599132623_331588.html

[10] Article extracted from h https://www.telecinco.es/informativos/sociedad/detienen-organizacion-criminal-explotacion-sexual-mujeres-mijas_18_2951295065.html

Negative pairs of the dataset—since we considered all the possible pairs of articles and most of them presented information about different cases.

Focusing on the importance of the features, we can observe that the model successfully distinguished which features were computing similarity and which were computing distance, as presented in *Figure 8: Feature importance of the pairwise classifier for detecting articles about the same case*.

The most relevant feature in the classification is the cosine similarity of TF-IDF. In *Example 3*, we can observe that it is the metric that varies the most between both examples. Even if this feature has a great influence in the prediction, the combination of features was selected in order to optimize algorithm's performance. The rest of the features provide additional support for the similarity assessment of both articles, considering separately the parts of an article: headline, lead (headline and subtitle), and body.

The trained model was used to predict the probability that each pair of articles presents information about the same case. The prediction was used as an approximation of the normalized similarity of the articles involved.

Given the pairwise similarity of all the pairs of articles of the dataset, we used a Hierarchical Clustering algorithm to group articles concerned with the same case, based on the pairwise distance—the detailed process and clustering parameters can be found in section *3.3.5 Clustering*.

The output of the clustering provided 289 groups of articles of distinct sizes. Although the results are not perfect, we can observe that those articles, that are grouped together, are certainly about the same case; meaning that the errors of the clustering are biased towards not grouping articles that should go together (FNs), rather than grouping articles that should not go together (FPs).

Example 4: Clustering results

Clustering example – Cluster id: 0 – Cluster size: 4

Article's headlines:

A[11]: The gangrape of a girl, by seven soldiers, shocks Colombia. (Original: *La violación de siete militares a una niña a la que secuestraron conmociona Colombia*).

[11]Article extracted from https://www.antena3.com/noticias/mundo/violacion-siete-militares-nina-que-secuestraron-conmociona-colombia_202006275ef72a20f9cab90001735fe7.html

B[12]: Investigating the rape of a twelve-year-old indigenous girl by seven soldiers in Colombia. (Original: *Investigan la violación de una niña indígena de doce años por parte de hasta siete militares en Colombia*).

C[13]: The Colombian Army reports a new violation by the military against an indigenous minor. (Original: *El Ejército de Colombia informa de una nueva violación de militares contra una menor indígena*).

D[14]: Colombia investigates eight soldiers for a new rape of a 15-year-old girl. (Original: *Colombia investiga a ocho militares por una nueva violación a una niña de 15 años*).

Articles belonging to the same cluster not grouped by the algorithm

E[15]: Seven soldiers admit they raped a 13-year-old indigenous girl in Colombia. (Original: *Siete militares admiten haber violado a una niña indígena de 13 años en Colombia*). → Cluster id: 259. Cluster size: 1

F[16]: The rapes of indigenous girls corner the Colombian Army: "They were drunk and they took them against their will". (Original: *Las violaciones a niñas indígenas acorralan al Ejército colombiano: "Estaban borrachos y se las llevaron a la fuerza"*) → Cluster id: 134. Cluster size: 1

Clustering example – Cluster id: 67 – Cluster size: 2

Article's headlines:

A[17]: The existence of a video with sexual content where Prince Andrés appears is revealed. (Original: *Revelan la existencia de un vídeo de contenido sexual donde aparece el Príncipe Andrés*).

B[18]: Epstein case: New York Prosecutor's Office asks to question Prince Andrés after knowing a key testimony (Original: *Caso Epstein: la Fiscalía de Nueva York pide interrogar al príncipe Andrés tras conocerse un testimonio clave*).

The collection does not contain any other article about this case.

Clustering example – Two singletons about the same case

Cluster id: 284

[12]Article extracted from https://www.telecinco.es/informativos/internacional/violacion-nina-indigena-colombia-militares-be5ma_18_2968620024.html

[13] Article extracted from https://www.telecinco.es/informativos/internacional/ejercito-colombia-violacion-militares-nina-indigena_18_2970720353.html

[14] Article extracted from https://www.lasexta.com/noticias/sociedad/colombia-investiga-ocho-militares-nueva-violacion-nina-anos_202007055f0237dc5a257f0001a04b91.html

[15] Article extracted from https://www.elmundo.es/internacional/2020/06/26/5ef5c379fdddff785e8b4646.html

[16] Article extracted from https://www.elmundo.es/internacional/2020/07/01/5efcc4b921efa04d2a8b46df.html

[17] Article extracted from https://www.abc.es/estilo/gente/abci-exempleada-jeffrey-epstein-detalla-supuesto-video-sexual-principe-andres-202007231455_noticia.html

[18] Article extracted from https://www.lasexta.com/noticias/internacional/caso-epstein-fiscalia-nueva-york-pide-interrogar-principe-andres_202006095edf7a2cdf27f50001998c47.html

A[19]: Eight years of prison were given to the rapist discovered in 'The Ana Rosa Program' who offered false jobs to his victims. (Original: *Ocho años para el violador descubierto en 'El Programa de Ana Rosa' que ofrecía trabajos falsos a sus víctimas*).

Cluster id: 80

A[20]: A sexual predator gets convicted thanks to an 'AR' reporter: "He said he would ask me to get naked". (Original: *Condenan a un depredador sexual gracias a una reportera de 'AR': "Dijo que me pediría que me desnudase"*).

Example 4: Clustering results shows a compilation of different situations:

- Cluster 0: A cluster where two of the articles were missed by the algorithm and were assigned as singletons. Articles left out clearly present information about the same case. We set a strict distance threshold—refer to section *3.3.5 Clustering* for more details—that prevented cluster 0 to be joined with clusters 259 and 134. Even if this type of mistakes are the most frequent ones in our clustering, a larger threshold would have outputted many FPs; we prefer to have true clusters with few articles left out than fake clusters misleading the posterior analysis.
- Cluster 67: A perfect clustering.
- Clusters 80 and 284: Two articles that were not grouped together but should belong to the same cluster. Both articles present information about the same case from different perspectives. This type of errors results from articles' intrinsic dissimilarities and are difficult to prevent.

The distribution of the sizes of the clusters resembles an exponential distribution: there are many clusters containing only one article and very few big clusters, meaning that there is a large difference in attention that any 2 cases will receive from the media—clusters' sizes are illustrated in *Figure 9: Distribution of the number of articles per sexual violence case* – obtained by clustering articles.

The number of articles covering the same sexual violence case can be used for inferring cases' representation. We considered that those cases represented by only one article have a low degree of representation, those explained in two or three articles have a medium level of coverage—since they have been covered by around 20% of the media outlets—and cases

[19] Article extracted from https://www.elmundo.es/espana/2020/07/30/5f21e368fc6c833a218b45fb.html

[20] Article extracted from https://www.20minutos.es/noticia/4340150/0/gracias-a-una-reportera-de-ar-condenan-a-un-depredador-sexual-dijo-que-me-pediria-que-me-desnudase/

covered by more than three articles can be considered highly represented—they have been shared by more than 30% of the outlets. According to this classification, only a reduced number of cases (less than the 10%) have a high level of exposure, while most cases are explained by only one article.

Media outlets are able to decide which are the 'privileged' cases that are worthy to be published and broadcasted, generating the large differences in coverage that we found in the dataset. For example, the biggest case from *Example 4: Clustering results* was shared in 6 articles; two of them were published by the same media outlet, El Mundo. In contrast, there are only two articles about the case presented by cluster 67.

Therefore, it is extremely important to detect which sexual violence cases must have more representation in the media.

5.3 Articles' analysis

Articles' analysis was divided into two main parts: The first one compares the representation of the cases in our dataset with official statistics, in order to detect which characteristics of sexual violence cases are required to be more represented in the media; The second one was based on the extraction of the types of information presented in sexual violence articles, used to study their relationship with popular stigmas related to the concerning topic.

It is important to remember that the results of this part were extracted by means of regular expressions and self-developed rules.

5.3.1 Coverage of cases

We analyzed the coverage of cases by inferring three characteristics of each case: the type of sexual violence, the bond between the victim and the perpetrator and the place where the crime took place. Each of them was compared with legitimate information as it can be found in section *2.1 Sexual violence statistics in* Spain.

Our main concern regarding these statistics arises from the nature of sexually violent crimes. Official statistics' results depend hugely on the circumstances of the primary information collected. Each case should be understood as a way of capturing reality, without being considered 100% true. Figures provided from INE show the number of people convicted for sexual violent crimes. However, most of the cases are not even reported, while the majority of the reported ones end up in agreements between the people involved. Therefore, sexual violent offences which satisfy these conditions cannot be captured by INE's statistics.

Similarly, *Macroenceusta 2019* presents other type of flaws, mainly, due to its general focus on violence against women.

No previous study has investigated sexual violence in Spain, exclusively, which points towards the existence of a deep knowledge gap in this field of study.

In our study, we focused on two coverage perspectives:

- Case selection: determining the type of cases that are considered more "worthy" of being published by media outlets. It can be assessed through cases' coverage.
- Case broadcasting: determining the type of cases that are widely spread by media outlets. It can be assessed through clusters' characteristics.

Type of sexual violence

We contrasted the percentage of cases of each type of sexual violence crime with the proportion of people convicted during 2019 (Instituto Nacional de Estadística, 2020), finding out that sexual abuse and sexual assault cases of our dataset resemble with a good precision the statistics considered—refer to *Table 19: Number of cases and their corresponding articles per type of sexual violence.*

Sexual harassment is the only type of sexual violence that presents significant differences in the comparison. Results denote that media outlets tend to disregard sexual harassment crimes: given 5 harassment crimes, less than 2 of them will be covered in any article. Generally, they do not receive much attention from the media, with the exception of a really low number of cases.

Example 5: Largest sexual harassment cluster in the dataset

Largest sexual harassment cluster in the dataset

Cluster id: 40; Cluster size: 6

Articles' headlines:

A[21]: David Villa denies the allegations after being investigated for sexual harassment. (Original: *David Villa niega las acusaciones tras ser investigado por acoso sexual*). – Media outlet: OK Diario

[21] Article extracted from https://okdiario.com/deportes/david-villa-niega-acusaciones-ser-investigado-acoso-sexual-5932249

B[22]: A New York City intern accuses Villa: "He masturbated me every day and my bosses laughed" (Original: *Una becaria del New York City acusa a Villa: "Me tocaba todos los días y mis jefes se reían"*). – Media outlet: El Confidencial

C[23]: David Villa, investigated in New York for alleged sexual harassment. (Original: *David Villa, investigado en Nueva York por presunto acoso sexual*) – Media outlet: El Mundo

D[24]: David Villa, investigated by the New York City for sexual harassment accusations (Original: *David Villa, investigado por el New York City por acusaciones de acoso sexual*). – Media outlet: A3 Noticias

E[25]: David Villa, investigated for alleged sexual harassment (Original: *David Villa, investigado por presunto acoso sexual*) – Media outlet: El Periódico

F[26]: David Villa, investigated for alleged sexual harassment (Original: *David Villa, investigado por presunto acoso sexual*) – Media outlet: ABC

The most mediatized sexual harassment case in the dataset is shown in *Example 5: Largest sexual harassment cluster in the dataset*, which supports our findings about the coverage of sexual harassment cases: the media generally turns their backs on these type of sexual violence, unless the case has some exceptional qualities, in the case of the example, a renowned Spanish football player who was accused of sexual harassment.

Observing outlet's behavior with respect to sexual harassment crimes, we can only agree with Walton's (2020) conclusions about media's role in the perpetration of sexual harassment stereotypes: "[sexual violence's victims'] experiences are downplayed, retaliation was commonly referenced as either an experience or something they feared, and they can find support in other victims." (Walton, 2020).

[22] Article extracted from https://www.elconfidencial.com/deportes/futbol/2020-07-23/david-villa-acoso-sexual-new-york-city-nueva-york_2693039/

[23] Article extracted from https://www.elmundo.es/deportes/futbol/2020/07/23/5f18bb5921efa059268b465b.html

[24] Article extracted from https://www.antena3.com/noticias/deportes/futbol/david-villa-investigado-new-york-city-acusaciones-acoso-sexual_202007235f19416f8fbe650001d6bc91.html

[25] Article extracted from https://www.elperiodico.com/es/deportes/20200723/david-villa-acoso-sexual-futbol-acusacion-barca-8051188

[26] Article extracted from https://www.abc.es/deportes/futbol/abci-david-villa-investigado-presunto-acoso-sexual-202007230145_noticia.html

The bond between the victim and the perpetrator

We considered four different types of bonds between the victim and the perpetrator and none of them is represented in the collection according to the statistics provided by the *Macroencuesta* 2019—included in *Table 2: Prevalence of different type of bonds between the victim and the perpetrator of sexual violenc*. These official statistics detail that the most common type of bond between the victim and the criminal is in decreasing order: "relationship", followed by "acquaintance", "stranger" and "relative". However, media outlets seem to represent a totally different reality since the most frequent type of crime is the least broadcasted one.

The opposite happens with the remaining types of bonds—refer to *Table 21: Number of cases and their corresponding articles per type of bond between the victim and the perpetrator* for detailed results—The most popular and mediatized sexual violence cases are those characterized by no prior interaction between the victim and the criminal; this type of bond is overrepresented by a factor of 4, followed by "acquaintance" and "relatives" bonds which have a disproportionate coverage too.

This misrepresentation of the reality goes along with preconceived notions and myths about sexual violence as perpetrators have always been demonized by means of monsters imagery in news articles (O'Hara, 2012).

One major issue in this debate is that sexual violence crimes that occur in the privacy of a relationship are hardly reported publicly, especially if it involves sharing a vulnerable situation. However, relationship-type bond cases that do appear in the dataset were not significantly broadcasted either. We must not forget that the lack of representation of sexual violence crimes caused by victim's trusted significant other feedbacks the myths and stereotypes of these abundant and undesired crimes.

Nonetheless, if we do not consider relationship-type bonds, we still obtain a skewed representation of reality by the over-representation of sexual violence crimes where the victim and the perpetrator were strangers.

Place where the crime occurred

For analyzing the place where perpetrators committed the crime, we assessed the most likely type of place from all the articles grouped in one case. It is important to remember that the analysis of places does not take into account those crimes in which the bond between the victim and the perpetrator was "relationship".

The categories considered, sorted in decreasing percentage order according to the *Macroencuesta* 2019, are house, public place, leisure space, workplace and educational place. We found out that only public places are fairly represented in the data collection.

As it can be seen in *Place of the crime* subsection of chapter *4 RESULTS*, the under-represented types of places are house, workplace and education. In the case of the workplaces, their articles may be discarded by media outlets as a precaution action in order to avoid possible lawsuits from companies; indeed, none of the crimes that happened in a workplace were broadcasted by more than 2 articles. As for sexual violence offenses that occurred in educational places, they are really infrequent in the dataset; nonetheless, most of them have large cluster sizes.

Finally, there is a clear relationship between the under-representation of crimes happening in houses and the over-broadcasting of sexual violence in leisure places. Results keep reiterating that media outlets report 'extreme' cases, those that will get people to talk, and those are, precisely all the articles that feedback pre-conceptions, myths and stereotypes reported in section *2.2 Representation of sexual violence in the media*.

As Friedman (2007) states "The social responsibility of business is to increase its profits", since media outlets are companies, by definition they want to generate profit and this is why they chose extreme cases; while reporters may care about reporting the actual facts, executives and stakeholders care about the bottom line.

5.3.2 Content analysis

The analysis of the content was done by considering each article individually (without taking into account cases aggrupation) and studying the presence of different types of information.

It is important to highlight that even if a large part of the aspects examined in this section coincide with the ones explored in the coverage analysis, the application of different methods, provided divergent results, emphasizing approaches' relevance when interpreting results.

The main goal of this content analysis is to obtain insights into the nature of information we are concerned with, typically presented in articles deemed to describe sexual violence. We focus on (1) information about the offence, (2) the presence of stigmas and stereotypes, and (3) expression indicating doubt about the information presented and euphemisms.

The main results considered in this part are *Table 25: Content analysis results – presence of features in articles' headline, subtitle and body* and the most relevant association rules found in the dataset described in *Table 26: Top-15 case-related association rules sorted by lift*, *Table*

27: Stigma and expression-related association rules and *Table 28: Top-15 inter-categories association rules sorted by lift*.

Sexual violence terms and case-related information

The presence of terms in each part of the article (considering article's headline, subtitle and body) indicates the relevance that the journalist attributes to that specific information (Conboy, 2007). In the case of sexual violence-related terms, we observed that most articles include terms indicating sexual assault; interestingly, the proportion of cases dedicated to sexual assault is considerably smaller, denoting that those terms explaining sexual assault acts are frequently used in articles that describe abuse. E.g., words describing prostitution have been considered as sexual assault-related terms, therefore, the article displayed in *Example 6* contains information about assault while describing sexual abuse.

Example 6: Sexual assault-related term in an article describing sexual assault.

Article's headline: Two brothers arrested for prostituting two minors at their home in Barcelona (Original: *Detenidos dos hermanos por prostituir a dos menores en su domicilio de Barcelona*).

Source[27] media outlet: El Periódico.

Regarding the presence of sexual harassment terms, we can observe that journalists pay little attention to these sexual offences since they prevail in the articles' bodies, attached as additional information, rather than being emphasized in articles' leads. This behavior perpetrates the understatement of sexual harassment, which according to Walton (2020) is the most common stigma around harassment in the media (Walton, 2020).

Example 7: Sexual abuse case stating harassment-related information in articles' body

Article's headline: 21 skating coaches investigated in France for child abuse (Original: *21 entrenadores de patinaje artístico investigados en Francia por abusos a menores*).

A fragment from the body: [skating coaches] were accused of sexual harassment and assault, three of them had already been convicted of similar acts in the past. (Original: *[los entrenadores] fueron acusados de agresiones sexuales y acoso, tres de ellos habían sido condenados por actos similares*)

Source[28] media outlet: El Periódico.

[27] Article extracted from https://www.elperiodico.com/es/barcelona/20200903/detenidos-hermanos-prostituir-menores-barcelona-8097727

[28] Article extracted from https://www.elperiodico.com/es/deportes/20200910/abusos-entrenadores-patinaje-artistico-francia-fiscalia-abitbol-8107468

The offences accentuated, primarily in articles, seem to describe sexual violence as simply sexual abuse and assault (ignoring harassment crimes).

Turning to case-related information, we observe that peoples' ages are usually added in articles; however, it is difficult to discern whether it is a peculiar trait of articles portraying sexual violence offences or it is a characteristic satisfied by all sort of news articles. Contrarily, we obtained convincing figures about the little importance bonds between people and places have, as these are mainly present in the main body of an article (it is important to emphasize that we are not focusing on any specific bond nor on the place where the crime occurred, but rather on the presence of this type of information); with the exception of public spaces, which do appear often in headlines.

Associations rules regarding sexual violence acts and case-related information—detailed in *Table 26*—can be summarized by the following ideas:

- If terms related to relationships appear in an article, that article has a high probability of containing information about familial bonds. This association can be understood through the idea that journalists' main information source is the police, who obtain information about a relationship by interrogating people close to the couple, i.e., their family.
- The presence of terms related to relationships or familial bonds imply the existence of information about homes.

Stigmas and expressions

The analysis of stigmas has been carried out focusing on five popular preconceptions, studying the presence of (1) peoples' origin—information stating their nationality or ethnicity—(2) terms or expression indicating intoxication, (3) clothes or expressions describing what someone was wearing during the time the crime was committed, (4) words demonizing the perpetrator or describing him as a mentally disturbed person and (5) vulnerability stigmas about the victim.

The most common stigma found in the dataset is the one related to the vulnerability of the victim; moreover, it represents the highest proportion of appearances in headlines. The vulnerabilities considered in the present study are comprised of two opposite types of stigmas: on the one hand, the victims' promiscuity; and on the other hand, the idealization of victims. Therefore, these results can be interpreted as evidence of the presence of the numerous stigmas

surrounding victims in articles, and their perpetration enhanced by media outlets, without being able to discern which of the two prejudices considered is the most prevalent one.

The second most frequent stigma is the origin, which has been detected in more than half of the articles analyzed, mainly in their main bodies.

Generally, we found little presence of the rest of the stigmas considered. Interestingly, from all of them, the least frequent prejudices are those concerning the perpetrator. This fact reveals that media outlets are trying to stay away from the stereotypes that imply directly the people involved, however, those preconceptions that persist are focused on the victims.

Focusing on the analysis of expressions, we examined whether euphemisms and uncertain expressions appeared in articles. Surprisingly, only a reduced proportion of articles contained euphemisms, while over half of them contained terms reporting uncertainty about the events related.

Example 8: Example of expressions of doubt in articles' headlines.

Article's A headline: A minor, in critical condition after her alleged rapist set her on fire (Original: *Una menor, en estado crítico después de que su supuesto violador le prendiera fuego*).

Source[29] media outlet: Antena3.

Article's B headline: A man has been arrested for allegedly masturbating before minor girls in Dos Hermanas, Seville (Original: *Detenido por masturbarse presuntamente ante niñas menores en Dos Hermanas, Sevilla*).

Source[30] media outlet: Telecinco.

Journalists tend to use terms such as 'presumed' and 'alleged' for relating cases that have not been proved yet. However, this type of vocabulary, aggregated with the misrepresentation of the reality towards stigmatized cases end up composing detrimental articles about a sensitive and relevant topic.

Association rules highlighted the following connections between antecedents and consequents:

[29] Article extracted from https://www.antena3.com/noticias/mundo/menor-estado-critico-despues-que-supuesto-violador-prendiera-fuego_202006155ee79f271a23fe0001b4d2c5.html

[30] Article extracted from https://www.telecinco.es/informativos/sociedad/detenido-masturbarse-menores-dos-hermanas-sevilla-exhibicionismo_18_3011670208.html

- Those articles containing stigmas related to intoxication (or intoxication and vulnerability) are highly likely to mention someone's nationality or origin.
- The presence of euphemisms in an article implies, with almost 90% confidence, stigmas about victim's vulnerabilities.
- Articles containing terms indicating vulnerabilities and intoxication are clear antecedents for expressions denoting doubt. Meaning that little credibility is given to vulnerable people.

Finally, we also checked whether there were additional associations considering all the aspects studied, finding two relevant association:

- When an article presents stigmas about vulnerability and an acquaintance bond, it is highly likely that the same article contains doubtful expressions.
- The mention of a leisure space and stigmas about a victims' vulnerable state implies sexual abuses.

Articles' analysis reflects how stigmas are being perpetuated by media outlets, however, articles have shown that these myths are not only maintained through explicit vocabulary but also by the biased representation of the facts, which, by no means, is adjusted to the reality of any of the statistics considered.

6 CONCLUSIONS & FUTURE WORK

This study was designed for the analysis of sexual violence coverage in Spanish media, with the objectives of (1) creating a dataset of sexual violence news articles; (2) study the coverage of sexual violence for each type of sexual violence, the bond between the victim and the perpetrator and the place where the crime occurred; and (3) get insights on the general writing characteristics of news articles, focusing on the perpetration of stigmas, stereotypes, and myths.

The first conclusion drawn from the present study is that activity of media outlets in Twitter can be used as a proxy of their interests and, therefore, we can achieve a dataset that represents in a fair way a media outlet using Twitter as the primary source.

The coverage analysis of sexual violence demonstrated that media outlets prefer to publish extreme sexual violent cases to increase their profit rather than to offer a more balanced representation of the reality. This finding is supported by the under-representation of harassment cases found in the dataset, the virtually inexistent coverage of sexual violent crimes that take place in homes and the over-representation of offences that happen in leisure spaces, and the fact that sexual violence cases where perpetrators and victims were in a relationship are ignored by the media. With these actions, media outlets shift the attention towards stereotyped situations, promoting stigma perpetuation. Moreover, outlets only broadcast the most frequent types of sexual violence when they have a special component that ensures the popularity of the article (such as celebrities involved).

Analogously, content analysis showed that the information provided by articles deemed to describe sexual violence (without considering the type of case they covered) is biased towards the understatement of harassment offences and the emphasis on assault and abuse crimes. Moreover, the only case-related facts highlighted in the headings are those that satisfy the already existing myths' requirements. Focusing on prejudices, journalists have tried to reduce their bias in the article; however, those that prevail are mainly focused on the victims. Additionally, we have observed that the presence of vulnerable victims has a high association with the presence of expressions related to doubt.

All these finding support that the coverage of sexual violence in the Spanish media does not represent the reality of this highly sensitive topic, emphasizing and providing a harmful portrayal of this field, which has an effect on the way that the population perceives sexual violence.

Overall, we can state that this analysis is feasible, as it can be seen in section *4 RESULTS*.

However, we must take into consideration that a larger dataset would help us establish a greater degree of accuracy on this matter. Further research could be done by improving the present work and including new attributes and characteristics into the analysis. If the investigation is to be moved forward, I would recommend the following improvements over the present work:

- Instead of collecting a fixed number of tweets for each media outlet considered, collect all the tweets comprised in a specific date range. By doing so, further analysis on specific cases could be done.
- Improve the tweets' classification by selecting a more sophisticated approach.
- Consider a larger number of official statistics capturing similar information for obtaining deeper insights into the coverage of sexually violent crimes, by paying special attention to the flaws of official statistics (flaws which originate from the way information is collected by institutions of the State). Reliable and recommended sources of information usually capture few aspects about sexual violence; e.g., labor unions have the most trustworthy information about sexual violence in workplaces (Andres-Pueyo et al., 2020).

Recommendations on further analysis:

- Include more features in the content analysis such as the gender of the author.
- Having a greater dataset able to represent each media outlet, examine the differences between media outlets in terms of coverage and writing style.
- Perform a study considering the more mediatized cases (biggest clusters) and compare how the articles' tone, in terms of doubt and stigmas, evolves as new information is discovered about the case.

7 BIBLIOGRAPHY

Ahmad, A. N. (2010). Is Twitter a useful tool for journalists? *Journal of Media Practice*, *11*(2), 145–155. https://doi.org/10.1386/jmpr.11.2.145_1

Andres-Pueyo, A., Nguyen, T., Rayó, A., & Redondo, S. (2020). Análisis empírico integrado y estimación cuantitativa de los comportamientos sexuales violentos (no consentidos) en España. *Análisis Empírico Integrado y Estimación Cuantitativa de Los Comportamientos Sexuales Violentos (No Consentidos) En España*, 0–275.

Aroustamian, C. (2020). Time's up: Recognising sexual violence as a public policy issue: A qualitative content analysis of sexual violence cases and the media. *Aggression and Violent Behavior*, *50*, 101341. https://doi.org/10.1016/j.avb.2019.101341

Bahrdwaj, Saksham; Saxena, N. (2019). *Word Mover's Distance for Text Similarity*. Towards Data Science. https://towardsdatascience.com/word-movers-distance-for-text-similarity-7492aeca71b0

Basch, C. H., Hillyer, G. C., Erwin, Z. M.-, Mohlman, J., Cosgrove, A., & Quinones, N. (2020). News coverage of the COVID-19 pandemic: Missed opportunities to promote health sustaining behaviors. *Infection, Disease & Health*, *25*(3), 205–209. https://doi.org/10.1016/j.idh.2020.05.001

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. *Society for Industrial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics 130*, *1*, 243–254. https://doi.org/10.1137/1.9781611972788.22

Calvo, S. T., Cervi, L., Jaraba, G., & Tusa, F. E. (2020). Spanish journalists on Twitter: Diagnostic approach to what, and how Spanish journalists talk about politics, international affairs, society, communication and culture. *Analisi*, *63*, 1–18. https://doi.org/10.5565/REV/ANALISI.3333

Campello, R. J. G. B. (2007). A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, *28*(7), 833–841. https://doi.org/10.1016/j.patrec.2006.11.010

Cañete, J. (2019). *Compilation of Large Spanish Unannotated Corpora*. Zenodo. http://doi.org/10.5281/zenodo.3247731

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Perez, J. (2020). Spanish Pre-Trained BERT Model. *Workshop Paper at PML4DC, ICLR*, 1–10.

Chakraborty, T., Mukherjee, A., Rachapalli, S. R., & Saha, S. (2018). Stigma of sexual violence and women's decision to work. *World Development*, *103*, 226–238. https://doi.org/10.1016/j.worlddev.2017.10.031

Chen, H., Huang, X., & Li, Z. (2020). A content analysis of Chinese news coverage on COVID-

19 and tourism. *Current Issues in Tourism*, 1–8. https://doi.org/10.1080/13683500.2020.1763269

Chen, Y., Perozzi, B., Al-Rfou, R., & Skiena, S. (2013). *The Expressive Power of Word Embeddings*. *28*.

Conboy, M. (2007). The language of news. *Angewandte Chemie International Edition, 6(11), 951–952.*, 13.

Daher, M. (2003). World report on violence and health. *Journal Medical Libanais*, *51*(2), 59–63. https://doi.org/10.1007/bf03405037

Damstra, A. (2019). Disentangling Economic News Effects: The Impact of Tone, Uncertainty, and Issue on Public Opinion. *International Journal of Communication*, *13*(0), 20.

De Benedictis, S., Orgad, S., & Rottenberg, C. (2019). #MeToo, popular feminism and the news : A content analysis of UK newspaper coverage. *European Journal of Cultural Studies*, *22*(5–6), 718–738. https://doi.org/10.1177/1367549419856831

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

DiBennardo, R. A. (2018). Ideal Victims and Monstrous Offenders: How the News Media Represent Sexual Predators. *Socius: Sociological Research for a Dynamic World*, *4*, 237802311880251. https://doi.org/10.1177/2378023118802512

Evans, A. (2018). #MeToo: A Study on Sexual Assault as Reported in the New York Times. In *Occam's Razor* (Vol. 8, p. 2018). https://cedar.wwu.edu/orwwu/vol8/iss1/3

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

Flanders, C. E., Anderson, R. E., Tarasoff, L. A., & Robinson, M. (2019). Bisexual Stigma, Sexual Violence, and Sexual Health Among Bisexual and Other Plurisexual Women: A Cross-Sectional Survey Study. *The Journal of Sex Research*, *56*(9), 1115–1127. https://doi.org/10.1080/00224499.2018.1563042

Ghoshdastidar, D., Perrot, M., & Von Luxburg, U. (2018). Foundations of comparison-based hierarchical clustering. *ArXiv*, *NeurIPS*.

Gottfried, J., & Shearer, E. (2016). *News Use Across Social Media Platforms 2016*. http://assets.pewresearch.org/wp-content/uploads/sites/13/2016/05/PJ_2016.05.26_social-media-and-news_FINAL-1.pdf

Gupta, V. (2011). Named Entity Recognition for Punjabi Language Text Summarization.

International Journal of Computer Applications, *33*(3), 975–8887. http://www.learnpunjabi.org/pdf/ner for summarization.pdf

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

Hart, P. S., Chinn, S., & Soroka, S. (2020). Politicization and Polarization in COVID-19 News Coverage. *Science Communication*, *42*(5), 679–697. https://doi.org/10.1177/1075547020950735

Huang, A. (2008). Similarity measures for text document clustering. *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings*, 49–56.

Instituto Nacional de Estadística. (2020). *National results - convicted for sexual offences*. https://www.ine.es/dynt3/inebase/en/index.htm?padre=4746&capsel=4731

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *FastText.zip: Compressing text classification models*. 1–13. http://arxiv.org/abs/1612.03651

Kabir, M. M. J., Xu, S., Kang, B. H., & Zhao, Z. (2015). *A New Evolutionary Algorithm for Extracting a Reduced Set of Interesting Association Rules* (p. 99). https://doi.org/10.1007/978-3-319-26535-3_16

Kumar, N., Sonowal, S., & Nishant. (2020). Email Spam Detection Using Machine Learning Algorithms. *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 108–113. https://doi.org/10.1109/ICIRCA48905.2020.9183098

Kuyumcu, B., Aksakalli, C., & Delil, S. (2019). An automated new approach in fast text classification (fastText): A case study for Turkish text classification without pre-processing. *ACM International Conference Proceeding Series*, 1–4. https://doi.org/10.1145/3342827.3342828

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, *60*(230), 591. https://doi.org/10.1145/1772690.1772751

Leiva, V., & Freire, A. (2017). *Towards Suicide Prevention: Early Detection of Depression on Social Media* (pp. 428–436). https://doi.org/10.1007/978-3-319-70284-1_34

López, L. (2017). *Violencia sexual: la punta del iceberg de un problema invisible*. http://uvadoc.uva.es/bitstream/10324/13308/1/TFG-L893.pdf

Malik, N., Bilal, A., Ilyas, M., Razzaq, S., Maqbool, F., & Abbas, Q. (2020). *Plagiarism Detection Using Natural Language Processing Techniques*. *26*(1), 90–102.

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

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 DELITOS CONTRA LA LIBERTAD E INDEMNIDAD SEXUAL, 2019 anual.pdf

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

Murray, C., Crowe, A., & Akers, W. (2016). How Can We End the Stigma Surrounding Domestic and Sexual Violence? A Modified Delphi Study with National Advocacy Leaders. *Journal of Family Violence*, *31*(3), 271–287. https://doi.org/10.1007/s10896-015-9768-9

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *WIREs Data Mining and Knowledge Discovery*, *7*(6). https://doi.org/10.1002/widm.1219

Newman, N., Richard Fletcher, W., Schulz, A., Andı, S., & Kleis Nielsen, R. (2020). *Reuters Institute Digital News Report 2020*. Page 82.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. *Lecture Notes in Engineering and Computer Science*, *2202*, 380–384.

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

Ouyang, Y., & Waterman, R. W. (2020). Trump, Twitter, and the American Democracy. In *Trump, Twitter, and the American Democracy* (pp. 131–161). Springer International Publishing. https://doi.org/10.1007/978-3-030-44242-2_5

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*(1), 25–29. https://doi.org/10.5120/ijca2018917395

Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). *Semi - supervised learning : a brief review*. *7*, 81–85.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Smitha, N., & Bharath, R. (2020). Performance Comparison of Machine Learning Classifiers for Fake News Detection. *Proceedings of the 2nd International Conference on Inventive Research in Computing Applications, ICIRCA 2020*, 696–700. https://doi.org/10.1109/ICIRCA48905.2020.9183072

Sriram, S. (2020). An Evaluation of Text Representation Techniques for Fake News Detection

Using : TF-IDF, Word Embeddings, Sentence Embeddings with Linear Support Vector Machine. *Arrow@TU Dublin*. https://doi.org/10.21427/5519-h979

Stanford History Education Group, Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository*, 29. http://purl.stanford.edu/fv751yt5934

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Walter, D., Ophir, Y., & Jamieson, K. H. (2020). Russian Twitter Accounts and the Partisan Polarization of Vaccine Discourse, 2015–2017. *American Journal of Public Health*, *110*(5), 718–724. https://doi.org/10.2105/AJPH.2019.305564

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

Wei, Y. Q., Yang, R. H., & Liu, P. Y. (2009). An improved apriori algorithm for association rules of mining. *ITME2009 - Proceedings 2009 IEEE International Symposium on IT in Medicine and Education*, *3*(1), 942–946. https://doi.org/10.1109/ITIME.2009.5236211

Yogish, D., Manjunath, T. N., Yogish, H. K., & Hegadi, R. S. (2020). Ranking Top Similar Documents for User Query Based on Normalized Vector Cosine Similarity Model. *Journal of Computational and Theoretical Nanoscience*, *17*(9), 4531–4534. https://doi.org/10.1166/jctn.2020.9330

Yuan, X. (2017). *An improved Apriori algorithm for mining association rules*. 6. https://doi.org/10.1063/1.4977361

ANNEX 1. Data and code release

The dataset and code created for the present project can be found in the following GitHub repository: https://github.com/marilenabudan/spanish_media_coverage_sexual_violence.

Data release

The dataset was released in the 'data' directory of the repository in a zip file protected by password. The compressed file contains all the tweets (scraped, hydrated, and classified), and news articles collected during the development of the project; as well as the labeled data for cases clustering and the final results.

The password can be obtained by contacting me through email to the following email address: marilena.budan01@estudiant.upf.edu.

Code and resource

The code and resources follow the same structure as sections *3 METHODOLOGY* and *4 RESULTS*. They are split into three main parts:

- Dataset creation
- Cases' classification
- Articles analysis

All directories contain a README file detailing the content of each file and a directory named 'utilities' with the resources and additional code used.

ANNEX 2. Terms and regular expressions used for articles analysis

Tables below detail the terms and regular expressions used for each aspect analyzed in articles' analysis.

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'

C

ANNEX 3. Association rules

Association rules about general information. Thresholds:

- Minimum support: 0.10
- Minimum confidence: 0.50
- Minimum lift: 1.10

Antecedent		Consequent	Support	Confidence	Lift
Sexual abuse	Bond relationship	Bond relative	0.11	0.62	1.89
Place house	Bond relative	Bond relationship	0.11	0.50	1.80
Place public	Bond relationship	Bond relative	0.12	0.54	1.66
Sexual assault	Bond relationship	Bond relative	0.10	0.51	1.56
Place house	Sexual harassment	Bond relative	0.11	0.50	1.53
Sexual assault	Bond relationship	Place house	0.15	0.76	1.48
Place leisure	Bond relationship	Place house	0.18	0.75	1.47
Sexual abuse	Bond relationship	Place house	0.13	0.74	1.45
Age	Bond relationship	Place house	0.19	0.72	1.40
Place public	Bond relationship	Place house	0.16	0.71	1.38
Bond relationship		Place house	0.19	0.70	1.37
Place public	Bond relative	Place house	0.19	0.70	1.37
Place leisure	Bond relative	Place house	0.18	0.69	1.35
Bond relative	Sexual abuse	Place house	0.16	0.68	1.32
Bond relative	Sexual assault	Place house	0.15	0.67	1.31
Bond relative		Place house	0.22	0.66	1.29
Bond relative	Age	Place house	0.21	0.66	1.29
Place educational		Bond acquaintance	0.10	0.75	1.27
Bond relative	Sexual harassment	Sexual abuse	0.12	0.78	1.25
Time	Sexual abuse	Place public	0.18	0.95	1.22
Time	Sexual harassment	Place public	0.10	0.94	1.22
Sexual abuse	Sexual harassment	Place public	0.25	0.94	1.21
Place public	Place workplace	Place leisure	0.13	0.94	1.21
Sexual abuse	Sexual harassment	Bond acquaintance	0.19	0.70	1.20
Bond relationship	Sexual harassment	Sexual abuse	0.10	0.75	1.20
Place house	Time	Bond acquaintance	0.11	0.70	1.19

D

Bond relative	Sexual abuse	Place public	0.22	0.92	1.19
Place leisure	Sexual harassment	Bond acquaintance	0.22	0.70	1.19
Place public	Sexual harassment	Bond acquaintance	0.23	0.70	1.19
Place house	Sexual abuse	Bond acquaintance	0.23	0.70	1.19
Place leisure	Sexual abuse	Place public	0.45	0.92	1.18
Sexual assault	Sexual harassment	Sexual abuse	0.18	0.74	1.18
Sexual abuse		Place public	0.57	0.92	1.18
Time	Sexual abuse	Place house	0.12	0.60	1.18
Place house	Sexual abuse	Place public	0.30	0.92	1.18
Age	Sexual abuse	Place public	0.55	0.91	1.18
Age	Place workplace	Place leisure	0.15	0.91	1.18
Sexual assault	Bond relative	Sexual abuse	0.16	0.73	1.17
Place educational		Place public	0.12	0.91	1.17
Sexual abuse	Bond relationship	Place public	0.16	0.91	1.17
Place leisure	Sexual abuse	Bond acquaintance	0.34	0.69	1.17
Place workplace		Place leisure	0.16	0.91	1.17
Bond relative	Age	Sexual abuse	0.23	0.73	1.17
Place educational	Age	Place public	0.11	0.90	1.17
Bond relative		Sexual abuse	0.24	0.73	1.17
Place house	Bond relationship	Bond acquaintance	0.14	0.68	1.17
Sexual assault	Sexual abuse	Place public	0.41	0.90	1.16
Place public	Time	Place house	0.15	0.60	1.16
Time	Age	Sexual assault	0.23	0.84	1.16
Time		Place public	0.25	0.90	1.16
Time		Sexual assault	0.24	0.84	1.16
Bond relationship	Sexual harassment	Place public	0.12	0.90	1.16
Time	Age	Place public	0.24	0.90	1.15
Time	Sexual abuse	Bond acquaintance	0.13	0.68	1.15
Place leisure	Time	Bond acquaintance	0.16	0.68	1.15
Sexual assault	Place workplace	Place leisure	0.10	0.89	1.15
Place leisure	Time	Place house	0.14	0.59	1.15
Place leisure	Sexual harassment	Place house	0.19	0.58	1.14
Sexual abuse	Sexual harassment	Place house	0.16	0.58	1.14
Place house	Sexual harassment	Bond acquaintance	0.15	0.67	1.14

Sexual harassment		Bond acquaintance	0.27	0.67	1.14
Place public	Time	Bond acquaintance	0.17	0.67	1.14
Time	Sexual assault	Place public	0.21	0.88	1.13
Place leisure	Bond relative	Sexual abuse	0.18	0.71	1.13
Age	Sexual harassment	Bond acquaintance	0.26	0.66	1.13
Place workplace		Place house	0.10	0.58	1.13
Place leisure	Bond relative	Bond acquaintance	0.17	0.66	1.13
Time	Sexual assault	Bond acquaintance	0.16	0.66	1.12
Bond relative	Sexual harassment	Place public	0.14	0.87	1.12
Place public	Sexual abuse	Bond acquaintance	375.00	0.65	1.12
Time		Place house	0.16	0.57	1.12
Place public	Place leisure	Bond acquaintance	0.40	0.65	1.12
Place public	Sexual harassment	Place house	0.19	0.57	1.11
Place leisure	Bond relationship	Bond acquaintance	0.15	0.65	1.11
Sexual abuse		Bond acquaintance	0.41	0.65	1.11
Sexual assault	Sexual harassment	Bond acquaintance	0.16	0.65	1.10
Place leisure	Sexual assault	Place house	0.32	0.56	1.10
Place house	Place public	Bond acquaintance	0.27	0.65	1.10
Time	Age	Sexual abuse	0.19	0.69	1.10
Bond relative	Sexual harassment	Place leisure	0.13	0.86	1.10
Time	Age	Place house	0.15	0.56	1.10
Place house	Bond relative	Bond acquaintance	0.14	0.64	1.10
Bond relative	Sexual abuse	Bond acquaintance	0.15	0.64	1.10
Time		Sexual abuse	0.19	0.69	1.10
Age	Sexual abuse	Bond acquaintance	0.39	0.64	1.10
Time	Sexual assault	Sexual abuse	0.16	0.68	1.09
Place house	Place leisure	Bond acquaintance	0.27	0.64	1.09
Sexual assault	Bond relationship	Place leisure	0.17	0.85	1.09
Sexual assault	Sexual abuse	Bond acquaintance	0.29	0.64	1.09
Sexual assault	Sexual harassment	Place house	0.14	0.56	1.09
Sexual assault	Sexual abuse	Place house	0.25	0.56	1.08
Time		Bond acquaintance	0.18	0.64	1.08
Bond relationship		Place leisure	0.24	0.84	1.08
Age	Sexual harassment	Sexual abuse	0.26	0.68	1.08

F

Bond relative	Bond relationship	Place leisure	0.12	0.84	1.08
Place public	Sexual assault	Place house	0.31	0.55	1.08
Bond relationship	Sexual harassment	Place leisure	0.11	0.84	1.08
Place leisure	Sexual harassment	Sexual abuse	0.22	0.67	1.08
Place house	Age	Bond acquaintance	0.31	0.63	1.08
Place house	Time	Sexual assault	125.00	775.00	1.07
Place house		Bond acquaintance	0.32	0.63	1.07
Time	Age	Bond acquaintance	0.17	0.63	1.07
Sexual abuse	Bond relationship	Bond acquaintance	0.11	0.63	1.07
Place public	Sexual harassment	Place leisure	0.27	0.83	1.07
Age	Bond relationship	Place leisure	0.22	0.83	1.07
Place house	Age	Sexual abuse	0.32	0.67	1.07
Place leisure	Sexual abuse	Place house	0.27	0.55	1.07
Place public		Bond acquaintance	0.49	0.63	1.07
Place leisure	Time	Sexual abuse	0.15	0.67	1.07
Sexual assault	Sexual harassment	Place public	0.20	0.83	1.07
Place public	Place leisure	Place house	0.33	0.55	1.07
Place public	Sexual assault	Bond acquaintance	0.35	0.63	1.07
Age	Sexual harassment	Place public	0.32	0.83	1.07
Time	Sexual assault	Place leisure	0.20	0.83	1.07
Place public	Age	Bond acquaintance	0.47	0.62	1.06
Sexual assault	Bond acquaintance	Place leisure	0.33	0.83	1.06
Place public	Bond relationship	Place leisure	0.18	0.83	1.06
Sexual assault	Bond relative	Place leisure	0.18	0.83	1.06
Place public	Time	Place leisure	0.21	0.83	1.06
Place house	Age	Place public	0.40	0.82	1.06
Place house	Age	Sexual assault	0.37	0.76	1.06
Bond relative		Place public	0.27	0.82	1.06
Place leisure		Place house	0.42	0.54	1.06
Place leisure	Age	Place house	0.40	0.54	1.06
Place educational		Sexual assault	0.10	0.76	1.05
Sexual abuse	Bond relationship	Place leisure	0.15	0.82	1.05
Place house	Sexual assault	Bond acquaintance	0.24	0.62	1.05
Place public	Bond relative	Bond acquaintance	0.17	0.62	1.05

Sexual harassment		Sexual abuse	0.27	0.66	1.05
Age	Bond relationship	Sexual abuse	0.17	0.66	1.05

Association rules about stigmas and expressions

Antecedent		**Consequent**	**Support**	**Confidence**	**Lift**
Stigma intoxicated		Stigma origin	0.13	0.73	1.26
Stigma vulnerability	Stigma intoxicated	Stigma origin	0.11	0.70	1.23
Expression euphemism		Stigma vulnerability	0.14	0.88	1.16
Stigma intoxicated		Stigma vulnerability	0.16	0.84	1.20
Stigma vulnerability	Stigma intoxicated	Expression doubt	0.10	0.68	1.06

Association rules containing both content-related and sitgmas and expression characteristics. Thresholds:

- Minimum support: 0.30
- Minimum confidence: 0.50
- Minimum lift: 1.05

Antecedent		**Consequent**	**Support**	**Confidence**	**Lift**
Stigma origin	Sexual abuse	Place public	0.34	0.94	1.22
Expression doubt	Sexual abuse	Place public	0.37	0.91	1.17
Stigma vulnerability	Sexual abuse	Place public	0.48	0.90	1.16
Place public	Expression doubt	Bond acquaintance	0.34	0.67	1.14
Place leisure	Stigma vulnerability	Sexual abuse	0.43	0.71	1.14
Stigma vulnerability	Sexual assault	Sexual abuse	0.40	0.71	1.13
Place leisure	Expression doubt	Bond acquaintance	0.33	0.66	1.12
Place public	Stigma origin	Bond acquaintance	0.30	0.66	1.12
Stigma vulnerability		Sexual abuse	0.53	0.53	0.70
Stigma vulnerability	Age	Sexual abuse	0.53	0.69	1.11
Stigma vulnerability	Bond acquaintance	Expression doubt	0.31	0.70	1.10
Stigma vulnerability	Bond acquaintance	Place public	0.37	0.85	1.09
Expression doubt	Age	Bond acquaintance	0.39	0.64	1.08
Expression doubt		Bond acquaintance	0.41	0.41	0.63

H

Stigma origin	Stigma vulnerability	Place public	0.35	0.84	1.08
Place public	Sexual assault	Stigma vulnerability	0.46	0.82	1.08
Place public	Age	Stigma vulnerability	0.61	0.82	1.07
Expression doubt	Age	Stigma vulnerability	0.50	0.82	1.07
Stigma vulnerability	Sexual abuse	Bond acquaintance	0.33	0.63	1.07
Sexual assault	Age	Stigma vulnerability	0.56	0.81	1.07
Place leisure	Age	Stigma vulnerability	0.60	0.81	1.07
Stigma origin		Bond acquaintance	0.36	0.63	1.07
Stigma vulnerability	Bond acquaintance	Place leisure	0.36	0.83	1.07
Sexual assault	Expression doubt	Stigma vulnerability	0.37	0.81	1.06
Place public	Stigma vulnerability	Place house	0.33	0.54	1.06
Stigma origin	Age	Bond acquaintance	0.35	0.62	1.06
Stigma origin	Stigma vulnerability	Sexual assault	0.32	0.77	1.06
Stigma vulnerability	Sexual abuse	Expression doubt	0.36	0.68	1.06
Stigma vulnerability	Sexual assault	Place leisure	0.47	0.82	1.06
Place leisure	Stigma vulnerability	Place public	0.50	0.82	1.06
Place house	Age	Stigma vulnerability	0.39	0.80	1.06
Stigma vulnerability	Expression doubt	Place public	0.41	0.82	1.06
Sexual assault	Age	Stigma origin	0.42	0.60	1.05
Place leisure	Stigma vulnerability	Place house	0.33	0.54	1.05
Stigma origin	Stigma vulnerability	Age	0.42	1.00	1.05
Place public	Sexual assault	Stigma origin	0.33	0.60	1.05
Place leisure	Stigma origin	Place public	0.36	0.82	1.05