The Coverage of Sexual Violence in Spanish News Media

Marilena Budan,[1] **Carlos Castillo** [1,2]
[1] Universitat Pompeu Fabra (UPF)
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA)
marilena.budan01@alumni.upf.edu, chato@icrea.cat

Abstract

The present study analyzes news articles about sexual violence published by online news media in Spain. Our goal is to get insights about the way in which sexual violence is covered and described, with a focus on biases described by previous research. We begin by collecting about 120,000 messages on Twitter ("tweets") posted during 2020 by 13 of the most popular online news outlets in Spain. Next, we use a supervised classifier to detect tweets that are likely to contain links to news articles related to sexual violence, finding nearly 500 of them. We group these into clusters of articles that are likely to refer to the same event, and extract a series of descriptive elements using regular expressions. Finally, we compare these descriptors with official statistics about sexual violence in Spain. Our conclusions find biases that are well aligned with those described in the literature about the coverage of sexual violence in the news, indicating that this type of automated analysis can help uncover these biases. For instance, news media covers sexual assault cases much more often than sexual harassment cases, despite the latter being more frequent. More worryingly, crimes happening at home are under-represented in the media, and crimes happening in leisure spaces are over-represented. In general, rather than presenting a balanced view of different types of sexual violence, media outlets perpetuate and reinforce harmful preconceptions and myths.

1 Introduction

According to a large survey conducted in 2020 by the Government of Spain, more than two million women 16 or older have suffered sexual violence at least once in their lives (Ministerio de Igualdad de España 2020). Official statistics also indicate that these crimes appear to be on the rise, as in 2019 crimes against sexual freedom and indemnity in Spain increased by 11.16% compared to the previous year, reaching a total of 15,319 cases (Ministerio de Interior de España 2019). Further analysis exploring the types and circumstances of these crimes indicates that sexual abuse is the most predominant (57.67%), in comparison to sexual assault and sexual harassment; furthermore, the perpetrator is usually someone from the victim's inner circle, and offenses occur, mainly, in trusted places and environments (Ministerio de Interior de España 2019).

News media tends to spread prejudice, myths and stereotypes about sexual violence (De Benedictis, Orgad, and Rottenberg 2019; DiBennardo 2018; Evans 2018; O'Hara 2012; Walton 2020), something that is reflected not only within the contents, but also on the language used to describe sexual violence in the news (Aroustamian 2020; Conboy 2007). Media outlets often treat perpetrators as "monsters" or "predators" that are demonized and described as people who have mental illnesses, alcohol/drug abuse problems, or both (O'Hara 2012; Walton 2020). There is also a tendency to idealize victims, which is harmful because any departure from a typically white, middle-upper class, "model" victim who has no previous sexual experience and is attacked by a complete stranger casts doubts on her story (DiBennardo 2018). The language used in sexual violence-related articles helps to perpetuate myths and fallacies regarding these types of crimes (O'Hara 2012), and it is common to find euphemisms and confusing language (Aroustamian 2020), such as "stealing someone's virginity." These expressions shift away the attention from the sexual violence case and instead emphasize and reinforce stereotypes.

Our research uses news articles written for a wide audience and published by large news media organizations in Spain. The contributions of the present research include:

1. an automated classifier to find news articles in Spanish about sexual violence,
2. a collection of URLs pointing to news articles on sexual violence published by large news media organizations in Spain,
3. a methodology to cluster these news articles into groups of articles that are likely to refer to the same event,
4. a collection of regular expressions to infer relevant descriptions from the news articles, and
5. an analysis in which we compare the prevalence of different descriptors with official statistics on sexual violence.

Our findings confirm that media outlets over-represent some cases of sexual violence, to the detriment of others, and reinforce stereotypes and stigmas around sexual violence.

2 Related work

In this section, we briefly overview previous work on the representation of sexual violence in the news, and automated

content analysis of news articles.

2.1 Representation of Sexual Violence in the News

Awareness of sexual violence has been on the rise, and it is increasingly recognized as a serious, global public health concern (World Health Organization 2022). Crimes against freedom and sexual indemnity have serious consequences on a victim's mental health. Due to the essential function that news media play in modern culture, media outlets can manipulate and skew readers' ideas and beliefs (Fitzpatrick 2018; Morgan 2018), and have an undeniable impact on the popular understanding of sexual violence through the implications of the language used in the description of cases involving these offenses (Flanders et al. 2019; Murray, Crowe, and Akers 2016).

Misconceptions about sexual violence, usually referred to by researchers as "*rape myths*" can be defined as "prejudicial, stereotyped or false beliefs, prejudices or stereotypes about rape, rapists, and rape victims" (Burt 1980). These myths, for instance, blame victims for their clothing or behavior, especially if they "are women of color, knew their assailant, drank alcohol, [or] dressed provocatively" (DiBennardo 2018).

People who believe rape myths are more likely to disbelieve the victim or side with the perpetrator, they are also more likely to accept sexual violence, and do not consider a rape is a "real rape" if it does not involve force or violence from a stranger (Mason and Monckton-Smith 2008). Misleading representations of sexual violence in the media "decontextualize abuse, encourage racism, promote stereotypes of women [...], blame victims and excuse assailants" (Kitzinger 2004). Perpetrators are portrayed as "beasts" or "perverts," which results in sexual violence being perceived not as a social issue, but as the result of unavoidable, random acts of violence (O'Hara 2012). Rape myths lead to harmful consequences, such as bad policymaking and the defunding of programs that actually work (De Vel-Palumbo, Howarth, and Brewer 2019).

2.2 Automated Content Analysis of News Articles

Due to the influence and impact that news articles have in the shaping of public opinion and the collective consciousness (Fitzpatrick 2018; Morgan 2018; Wineburg and McGrew 2016), methods for automated content analysis of news have attracted considerable attention from researchers. A common approach is to perform automated content analysis using Natural Language Processing (NLP) techniques. This requires a prior transformation of the text, since most Machine Learning (ML) models used in NLP are designed to take vectors as inputs, and not texts.

There are multiples ways of representing information depending on the linguistic characteristics we want to emphasize, from *one-hot encoding*, to TF-IDF weighting, and *text embedding* techniques, which allow us to represent terms, sentences and documents as vectors preserving their semantic and syntactic information (see, e.g., Sriram 2020).

Data transcribed to machine-readable formats can be used to perform diverse studies on news' content, from the extraction of topics or the merging of near-duplicates, to high-level tasks including fake news detection (e.g., Smitha and Bharath 2020) and content coverage analysis (e.g., Hart, Chinn, and Soroka 2020).

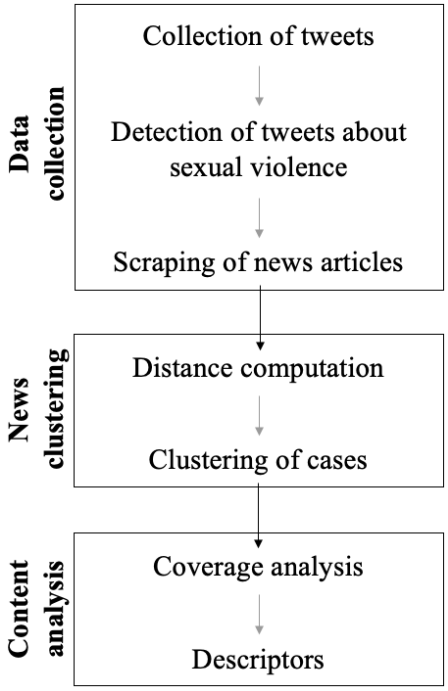

Figure 1: Data processing pipeline. The methodology has three main blocks: data collection, news clustering, and content analysis.

3 Data and Methods

This section describes how we collected, clustered, and annotated news articles. Figure 1 depicts our data processing pipeline.

3.1 Data Collection

We started with a list of the 15 generalist news media sources in Spain with the largest audience, according to Reuters Institute's Digital News Report (Newman et al. 2020). All of these use online social networking sites, such as Twitter and Facebook, to broadcast their content with the goal of increasing traffic to their websites (Ahmad 2010). On Twitter, these sources had multiple accounts, each dedicated to a different type of content. Having determined the most suitable account for the current goal, the most recent 10,000 tweets were collected on September 2020 from each of them, using the public API offered by Twitter.

Tweets were categorized as either being about sexual violence, or about some other topic, using a supervised classifier based on logistic regression. The supervised classifier was trained iteratively, starting with a classifier created by manually labeling the most recent 2,000 tweets of each source. Labeling was primarily done by one of the authors of this paper, who looked for any indication of sexual violence, which is defined through a list of possible offenses in

Media	Tweets		Articles
	All	On topic	
20 Minutos	10K	76	64
La Sexta	10K	78	45
El Confidencial	10K	39	18
RTVE News	10K	23	8
La Vanguardia	10K	15	13
ABC	10K	67	63
Cadena SER	10K	46	31
OK Diario	10K	31	11
El Periódico	10K	49	28
La Razón	10K	45	26
Telecinco	10K	103	94
El Mundo	10K	87	49
Antena 3	10K	65	46
Total	120K	883	496

Table 1: News sources used in this study. 10K tweets were downloaded from the main Twitter account of each news source and categorized automatically. "On-topic" indicates how many of them were about sexual violence according to our automatic classifier. We attempted to download all the news articles pointed by those tweets, but were able to download 50%-60% of them due to many cases of URLs pointing to section pages or home pages instead of article pages.

the official survey, and assigned the positive label to tweets mentioning any such offense, and the negative label to the remaining tweets (Ministerio de Interior de España 2019). Labeling criteria and edge cases were discussed among authors. Then, this classifier was applied to the entire corpus of downloaded tweets, and used to guide a larger manual labeling process, in which 5,000 additional tweets from each source were categorized. The resulting classifier has a precision of 85% and a recall of 83% (i.e., 17% false negative rate), as measured on a test set (30% of the labeled tweets).

Next, we considered all URLs contained in positively classified tweets that pointed to the domain corresponding to each Twitter account. Two of the original 15 sources (*El Pais* and *El Diario*) used a technique for hiding the URLs in the tweets, and hence were not considered. We followed the URLs and downloaded the target links, which was possible in 50%-60% of the cases due to many URLs pointing to section pages, home pages or to duplicate URLs instead of article pages. We converted downloaded articles to the NewsML-G2 form, a standardized way of storing articles in an XML format defined by the International Press Telecommunications Council (IPTC).[1]

Table 1 shows the number of tweets and articles collected at each step of the process. The final dataset used for the analysis contains 496 news articles.

3.2 News Clustering

A comparison with official statistics requires some level of aggregation beyond that of an individual news article. We found that it is fairly common that a single event is covered in multiple news articles within and across different news sources. Hence, we grouped these documents into clusters of articles describing the same event.

[1]https://iptc.org/standards/newsml-g2/

Category (Coverage)	Cluster size (Number of sources)	Number of clusters (Number of cases)
Low coverage	1	191
Medium coverage	2	51
	3	22
High coverage	4	8
	5	8
	6	3
	7	2
	8	3
	9	1
Total		289

Table 2: Sexual violence cases (clusters) classified by the number of sources covering them (clusters' size). Results from applying Hierarchical Clustering for grouping articles that probably describe the same sexual violent offenses into clusters. For instance, a singleton cluster (size = 1) is a sexual violence offense covered in only one source. A total of 496 articles are grouped into 289 cases.

To perform this clustering, we first build a supervised distance function. This distance function is simply one minus the probability that two articles refer to the same event. This probability is computed by a calibrated logistic regression classifier trained on a sample of a few thousand pairs of articles, manually labeled by the authors of this paper as either "same case" or "different case." The features used by the classifier are: (1) the Jaccard coefficient and Goodall metric between the named entities mentioned in the headlines and bodies of the articles (4 features), (2) the cosine similarity between the TF-IDF representation of the articles, (3) the distance between the BETO (a Spanish variant of BERT created by Cañete 2019) encodings, and the Earth Movers Distance between the headlines (2 features), (4) the difference in days between the publication of the articles, and (5) the difference in characters between the lengths of the articles. The resulting classifier has a precision of 94% and a recall of 90% (measured on a test set).

The clustering of the articles was created using Agglomerative Hierarchical Clustering, which groups objects to form a binary tree starting from singleton clusters, with the linkage modality of "average." Each resulting cluster gathers all articles describing the same sexual violent offense, i.e., a *case*. The number of different media sources present in each cluster suggest the amount of media coverage that media, in general, gave to each case. Table 2 shows the number of clusters obtained per cluster size; 66% of the offenses detected are covered by one source, while coverage of the remaining offenses (34%) fluctuates between 2 and 9 sources. We categorized cases covered by only one source as having *low coverage*, those covered by two or three sources as having *medium coverage*, and cases covered by more than three sources as *high coverage*. Less than 10% of the cases belong

to the high coverage category.

3.3 Automated Content Analysis

We evaluate news articles' content at two different levels: per-case (i.e., per cluster) and per-article.

Per-Cluster Analysis Analyzing the dataset at a cluster level, we can look at the media coverage and compare it with studies and official statistics on sexual violence. In particular, we look for discrepancies between the distributions of categories or descriptors in the cases covered in online news media in Spain, and the same distributions in surveys conducted by the Government of Spain. For this analysis, we consider three different attributes, described in sections 4.1 and 4.2:

- the type of sexual violence: sexual harassment, sexual abuse, or sexual assault;
- the bond described between the victim and the person who committed the crime: a romantic or sexual relationship, a familial bond, whether they were friends or acquaintances, or if they were complete strangers; and
- the place where the attack occurred: at home, in a public space, in a workplace, educational place, or in a place of leisure.

These attributes are determined through a series of regular expressions that are first applied at the level of individual articles, and then aggregated at the level of cluster (for instance, by majority class for the type of sexual violence). When evaluated over a small sample of 50 articles, these regular expressions achieved an F-measure (harmonic mean of precision and recall) between 0.75 and 0.95. The regular expressions, together with all our models and code, will be released with the camera-ready version of this paper.

Per-Article Analysis We also considered the presence or absence of specific types of information in different parts of news articles (such as headline, sub-headline, or body). The reason to consider these different parts is that they have different saliency, and indeed many readers, particularly those who find the article through social media, might read only the headline but not the body of an article. The information we focus on are stigmas that media usually carries related to victims and perpetrators, as well as expressions of doubt about the stories of the victims. These are also found through regular expressions. Further description of these categories can be found in Section 4.3.

4 Categories and Descriptors

In this section, we describe the categories of sexual violence analyzed, as well as descriptors of the circumstances in which attacks take place, and stigmas and expressions of doubt found in some news articles.

4.1 Types of Sexual Violence

This paper considered three different types of sexual violence, defined by the National Institute of Statistics (INE) of Spain as follows, with "sexual assault" being the most severe offense (Ministerio de Interior de España 2019).

- **Sexual harassment**: unwelcome sexual advances, requests for sexual favors, and other verbal or physical conducts of a sexual nature.
- **Sexual abuse**: an act of violence inflicted by the attacker against someone they perceive as weaker than them; a crime committed deliberately with the goal of controlling and humiliating the victim.
- **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 sexual acts.

4.2 Circumstantial Descriptors

Victim-perpetrator bond describes the relationship between the victim and the perpetrator, which, according to the literature, is usually one of acquaintance or previous or current partnership (Ministerio de Igualdad de España 2020):

- **Romantic and/or sexual relationship**: the victim and the perpetrator were in an affective relationship when the sexual violence occurred *or had been* in such a relationship before.
- **Familial/relative**: the victim and the perpetrator have a familial bond, i.e., they are relatives.
- **Acquaintance**: the victim and the perpetrator knew each other before the crime; this includes friends, teachers, colleagues, workmates, etc.
- **Stranger**: the victim and the perpetrator did not know each other.

Place of the crime describes the type of place where the attack took place. The survey shows that across type of bond (i.e., not limited to cases in which the victim and perpetrator are in a relationship), the most common type of space are homes (Ministerio de Igualdad de España 2020).

- **Public**: any public space such as streets, public squares, public transportation stations and bus stops, public transport vehicles, beaches, parks.
- **Workplace**: the place where someone works, such as an office, a shop, a coworking space, or a factory.
- **Educational**: a place where educational activities occur, such as schools, libraries, high schools, universities.
- **Leisure**: recreation places such as cinemas, theaters, bars, restaurants, shopping centers, nightclubs.
- **Home**: a place where people live, residences.

4.3 Stigmas and Expressions of Doubt

We considered various stigmas around sexual violence, gathered from the literature overviewed in Section 2.1. We consider five types of stigmas:

- **Origin**: whether information about the origin (nationality, ethnicity) of the victim and/or perpetrator is present in the article.

- **Intoxicated**: whether the article contains terms related to intoxication, such as terms related to alcohol or other drugs.
- **Clothing**: whether the article describes the clothes worn by the victim or perpetrator.
- **Vulnerability**: whether articles contain terms referring to someone in a vulnerable state, such as alone, young, minor, abandoned.
- **Aggressor**: whether articles contain stigmas commonly associated to perpetrators of sexual violence, such as mental illness, characterizing someone as a predator, and so on.

We also considered two types of expressions around the reliability of the account or the way in which the attack is described:

- **Doubt**: expressions casting doubt, showing incredulity, or hedging expressions, such as "alleged" or "presumed crime," among others.
- **Euphemism**: expressions that change or soften the severity of sexual violence, such as "stealing the innocence."

5 Results

5.1 Per-cluster categories and descriptors

Sexual violence type Table 3 compares the incidence of each type of sexual violence offenses in our dataset versus the statistics provided by Ministerio de Interior de España 2019. While the fraction of sexual abuse and sexual assault cases captured in the dataset resemble, to some extent, official statistics, harassment cases do not receive much attention from the media, as only 8% of the sexual violence cases in the news describe harassment situations, while these have a 22% incidence in the official survey.

Sexual violence type	Survey	News
Harassment	22%	8%
Abuse	58%	63%
Assault	20%	25%
Not classified		4%

Table 3: Coverage analysis of the types of sexual violence. Incidence (Survey) vs coverage (News) compares the percentage of cases in official sources (Ministerio de Interior de España 2019) against the percentage of cases according to news clusters or "cases"; columns add up to 100%.

Victim-perpetrator bond News articles seem to present an entirely wrong picture about the bond between the victim and the perpetrator, as we can see on Table 4. The *most* frequent type of relationship in actual sexual violence attacks, a current or past romantic and/or sexual relationship, is the *least* present one in the news. Conversely, one of the least frequent type of relationship (being complete strangers), is the most present one.

Victim-perpetrator bond	Survey	News
Romantic or sexual relationship	71%	11%
Familial/relative	6%	13%
Acquaintance	14%	28%
Stranger	11%	43%
Not classified		5%

Table 4: Coverage analysis of the types of bond between the victim and the perpetrator. Incidence (Survey) vs coverage (News) compares the percentage of cases in official sources (Ministerio de Igualdad de España 2020) against the percentage of cases according to news clusters or "cases"; columns add up to 100%.

Place where the attack occurs The survey we use for reference asks about the place of an attack only in cases in which victim and perpetrator did *not* have a romantic or sexual relationship. Hence, we remove the 11% of news in which this is the case before computing the statistics. Table 5 shows that most sexual violence happens at home, but this is severely under-represented in the media. Instead, media over-represents attacks in leisure spaces. Additionally, the proportion of cases in our dataset describing attacks in workplaces or educational places does not even reach half of their true incidence.

Place where the attack occurred	Survey	News
Home	65%	10%
Public space	36%	44%
Leisure place	16%	34%
Work place	7%	2%
Educational place	6%	1%
Not classified		9%

Table 5: Coverage analysis of the types of the places where sexual violence attacks happen. Incidence (Survey) vs coverage (News) compares the percentage of cases in official sources (Ministerio de Igualdad de España 2020) against the percentage of cases according to news clusters or "cases". Cases in which victim and attacker have/had a sexual/romantic relationship are not considered, to make the news data comparable with the survey; survey column adds up more than 100% as multiple answers were allowed in the survey, news column adds up to 100%.

5.2 Per-article expression of stigmas and doubts

Mentions of stigmas The presence in articles of information linked to stigmas around sexual violence is displayed in Table 6, as well as the parts of the article where they were found. Expressions linked to the vulnerability of the victim are the most common ones, being present in 76% of the articles and having a large presence in articles' headline; meaning that it is something authors tend to highlight. This is followed by the origin of perpetrator and/or victim, and references to intoxication and clothing, usually of the victim. Overall, explicit terms that can be associated with

preconceptions around sexual violence are avoided by authors. Even though, those that persist are more focused on the victims.

Expressions of doubt and euphemisms Hedging expressions or expressions of doubt (present in 64% of the articles) are more frequent than euphemisms (present in 16% of the articles). Both types of expression are more common in the body of the article than in its headline or sub-headline.

	Articles		Article element		
	Total	**%**	**Headl.**	**Subhead.**	**Body**
Origin	283	57%	10%	11%	57%
Intoxicated	92	19%	4%	2%	18%
Clothing	62	13%	1%	1%	13%
Perpetrator	20	4%	1%	0%	4%
Vulnerability	378	76%	32%	29%	76%
Euphemisms	77	16%	0%	1%	15%
Doubt	317	64%	14%	14%	62%

Table 6: Expressions in article texts that are linked to sexual violence stigmas or express doubts. The right portion of the table describes the part(s) of the article where the expressions are found.

	Media impact		
Type of sexual violence	**High**	**Medium**	**Low**
Harassment	25%	21%	30%
Abuse	12%	8%	1%
Assault	63%	71%	69%
Victim-perpetrator bond	**High**	**Medium**	**Low**
Romantic/sexual relationship	15%	30%	55%
Familial/relative	0%	28%	72%
Acquaintance	12%	26%	62%
Stranger	8%	22%	70%
Place	**High**	**Medium**	**Low**
Home	0%	36%	64%
Public space	9%	26%	65%
Leisure place	8%	23%	69%
Work place	0%	20%	80%
Educational place	67%	33%	0%

Table 7: Coverage across news sources analysis. Clusters belonging to each category are distributed according to the number of news sources covering each story: high impact (>3 sources), medium impact (2-3 sources), and low impact (1 source); rows add up to about 100%.

5.3 Coverage across news sources

Table 7 presents the impact of different categories of news cases, measured as the number of different media sources covering a case (high: more than 3, medium 2-3, low: 1). Only a few cases, less than 10%, have high impact, and most cases have low impact. Results have to be read as the media impact *when cases are covered*, which has been evaluated in Section 5.1. Cases that tend to have a relatively high impact involve victim-perpetrator that are or had been in a romantic/sexual relationship, and/or attacks that happen in public spaces or leisure places. Another interesting finding is that all sexual violent offenses happening in educational places have a high or medium impact when they are exposed by media outlets.

6 Conclusions

In general, results are aligned with previous findings in the literature. Media outlets decide which are the cases that are worthy to be published, generating the large differences in coverage that we find in the dataset. For instance, harassment cases are downplayed in the news, unless they involve renowned people (Walton 2020). In our data, attacks where the victim and the perpetrator had a romantic or sexual relationship, or attacks that happen at home, or in a place of work or education, are all covered less than what would correspond considering how frequent they are. In contrast, cases in which the perpetrator is a stranger, or that happen in a leisure place, are over-represented in the news. This finding is aligned with the surveyed literature on coverage of sexual violence in news media, as well as with a survey about perceptions of sexual violence in Spain, in which responders also believed that women were more likely to be assaulted by a stranger (Delegación del Gobierno para la Violencia de Género 2018).

News articles also tend to contain expressions of doubt and euphemisms. This agrees with a harmful trend noted by previous work, in which the media in general tends to downplay the experiences of sexual violence victims (Walton 2020). Furthermore, preconceived notions surrounding perpetrators' characteristics are reinforced by a skewed representation of the type of bond between the victim and the perpetrator, as media presents them, mainly, as strangers. Generally, we found some presence of stigmas in news articles, as, for instance, a minority of them includes references to intoxication (19%) or clothing (13%). Anecdotally, the least frequent prejudices we found were those concerning the perpetrator, however we did find these preconceptions focused on the victims. The usage of hedging expressions such as "presumed" and "alleged," could be justified if we consider that in most cases, by the time an article is published, the official investigation has just started. However, this type of vocabulary, compounded with the misrepresentation of the circumstances of different types of sexual violence, presents an unfavorable view of how news media in Spain treats this sensitive and relevant topic.

Limitations. Our comparison with official statistics assumes that the number of clusters of news articles of a certain type, can be used as a proxy of the extent to which that type is covered in the news. Naturally, the extent to which this is correct depends on many factors, including the accuracy of the classification method employed, and the extent to which one case corresponds to one cluster. For classification accuracy, we are using regular expressions and supervised classification methods that can be improved. Official statistics' results depend considerably on the circumstances of the primary in-

formation collected. Each case should be understood as a way of approximating a reality. Figures from the National Institute of Statistics of Spain show the number of people convicted for sexual violent crimes. However, most of the cases are not even reported, while many of the reported ones end up in agreements between the people involved. Furthermore, despite our efforts, the collection of articles is most likely incomplete. The number of articles obtained per media source is different, and we cannot ensure a perfect representation of the overall coverage of this topic in the Spanish news media.

Ethical issues. Regarding *personal data,* this research used exclusively news articles intended for a wide audience and published in high-visibility online news outlets in Spain; we did not attempt to gather any additional information about the cases described in them. Regarding *the well-being of the research team,* team members were asked not to engage with this research if they were triggered by news articles about sexual violence, were reminded at each of our weekly meetings to check themselves for signals of discomfort with the texts being examined, and had the option to perform research tasks not involving these texts at various points during the research.

Data availability. Our data release includes links to the tweets and links to the news articles, along with our labels. It also contains the keywords and regular expressions used to identify various content descriptors in the articles, plus a link to a longer report. https://github.com/marilenabudan/spanish_media_coverage_sexual_violence

Acknowledgments. This work has been partially supported by: the HUMAINT program (Human Behavior and Machine Intelligence), Joint Research Center, European Commission; "la Caixa" Foundation (ID 100010434), under the agreement LCF/PR/PR16/51110009; and the EU-funded "SoBigData++" project, under Grant Agreement 871042.

References

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

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.

Burt, M. 1980. Cultural myths and supports for rape. *Journal of Personality and Social Psychology* 38(2): 217–230.

Cañete, J. 2019. Compilation of large spanish unannotated corpora(May 2019). https://zenodo.org/record/3247731.

Conboy, M. 2007. *The language of the news*. Routledge London.

De Benedictis, S.; Orgad, S.; and 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.

De Vel-Palumbo, M.; Howarth, L.; and Brewer, M. B. 2019. 'Once a sex offender always a sex offender'? Essentialism and attitudes towards criminal justice policy. *Psychology, Crime & Law* 25(5): 421–439.

Delegación del Gobierno para la Violencia de Género. 2018. *Percepción Social de la Violencia Sexual*. Government of Spain. ISBN 978-84-7670-735-7.

DiBennardo, R. A. 2018. Ideal victims and monstrous offenders: How the news media represent sexual predators. *Socius* 4: 2378023118802512.

Evans, A. 2018. # MeToo: A study on sexual assault as reported in the New York Times. *Occam's Razor* 8(1): 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: 45–62.

Flanders, C. E.; Anderson, R. E.; Tarasoff, L. A.; and 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* .

Hart, P. S.; Chinn, S.; and Soroka, S. 2020. Politicization and polarization in COVID-19 news coverage. *Science Communication* 42(5): 679–697.

Kitzinger, J. 2004. *Media Coverage of Sexual Violence Against Women and Children*, 13–38. Blackwell.

Mason, P.; and Monckton-Smith, J. 2008. Conflation, collocation and confusion: British press coverage of the sexual murder of women. *Journalism* 9(6): 691–710.

Ministerio de Igualdad de España. 2020. Macroencuesta de Violencia contra la Mujer 2019. https://violenciagenero.igualdad.gob.es/violenciaEnCifras/macroencuesta2015/Macroencuesta2019/home.html.

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:34be8e1f-e3a5-42d3-a6e9-1a38e13e5598/Informe

Morgan, S. 2018. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy* 3(1): 39–43.

Murray, C.; Crowe, A.; and 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.

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

O'Hara, S. 2012. Monsters, playboys, virgins and whores: Rape myths in the news media's coverage of sexual violence. *Language and literature* 21(3): 247–259.

Smitha, N.; and Bharath, R. 2020. Performance comparison of machine learning classifiers for fake news detection. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, 696–700. IEEE.

Sriram, S. 2020. An evaluation of text representation techniques for fake news detection using: TF-IDF, word embeddings, sentence embeddings with linear Support Vector Machines .

Walton, N. 2020. *Myths, messaging, and the media: the media's role in perpetuating sexual harrassment stereotypes*. Ph.D. thesis, University of Delaware.

Wineburg, S.; and McGrew, S. 2016. Evaluating information: The cornerstone of civic online reasoning .

World Health Organization. 2022. Sexual violence. https://www.who.int/reproductivehealth/topics/violence/sexual_violence/en/.