Research Note

Natural Language Processing, Fall 2024, Hertie School

Instructor: Dr. Sascha Göbel

NLP: Classifying Gender-Based Violence Frames in News Reporting

Aranxa Márquez Ampudia (231357)

Github Repository for the Code

The primary aim of this research is to explore if gender-based violence (GBV) news articles can be classified into Episodic[1] and Thematic[2] categories using advanced natural language processing (NLP) techniques. This was done by building upon a labeled dataset that previously categorized over 900 news articles (Márquez Ampudia,2023). The current work gives a brief overview on the theory and real-life importance of the gender-based violence problem and its narrative in the media; a description of the data used to train the model, and it extends the study by employing TF-IDF Vectorization and Logistic Regression for a simple binary classification.

Framing theory and its importance in the coverage of gender-based violence

Framing theory is a way in which people simplify certain conflicts —or in this case, a social problem— by relating them to more familiar issues (Scheufele and Tewksbury, 2007). This is why the study compares the presentation of news at a macro level[3] to contrast the differences between local media on the issue of gender-based violence.

Like stereotypes, frames serve to interpret information and make sense of our experiences according to mental schemas about the world around us (Goffman, 1974). Framing a narrative can be done consciously or unconsciously both by the communication professionals and by those who receive the message; by the culture or even by the text itself. Each naturally tends to select and highlight features that construct arguments about issues, their possible causes and how they are evaluated or solved (Entman, R. 1993).

[1] News mentioning direct violence towards women or girls.

[2] Cultural or public policy-related articles portraying women or girls)

[3] The macro construct refers to the way journalists present information in such a way that it is related to 'pre-existing and underlying schemas in audiences' (Scheufele and Tewksbury, 2007, p.12).

A hypothetical example would be if the gender violence topic was mentioned mostly linked to feminicides. With this information, certain people in the audience would be more likely to associate these concepts with each other, omitting that there are more types of gender violence. On the other hand, the community structure theory and framing theory recognize the social and cultural influence as a whole force that affects the way in which journalists present certain issues in the news (Van Gorp, 2007). In that sense, as well as impacting on the way an issue is understood, it can have an impact on the actions that are taken to address it (Kahneman and Tversky, 1984).

Frames of gender-based violence

Due to the importance in the formation of the discourse on women victims of gender violence, several empirical studies have undertaken the task of contrasting the different ways in which this information is treated (Ulloa Luna and Spiller, 2014; Marín, Armentia and Caminos, 2011). The results of these studies empirically demonstrate that the issue of gender-based violence has gradually shifted from episodic and violent events, usually with yellowish tones found in the sections of news reports, to refer to it as a structural problem with serious repercussions on the care and protection of women's human rights.

Another element that stands out in studies on the coverage of gender-based violence is the identification of actors, as Colombini et al (2016) also did, although in terms of public policy. Angélico, Dikenstein, Fischberg and Maffeo (2014) observe, for example, to what extent the voice of the victim is rescued when reporting violent events. Although it is difficult to recover the victim's voice in the case of femicides or gender-related killings of women, it is possible to do so anecdotally through people close to the victim, such as family members. However, the results of Angélico et al. reveal that it is the voices of the perpetrators and the people in the judicial sphere who have more echo on the media podium (p. 285).

Methodology

In accordance with the types of media coverage that have been documented, two major mutually exclusive frames were defined for the main analysis: the *Episodic* and the *Thematic* (Márquez Ampudia, 2023; Escribano González, 2014).

For this research note the focus will remain only in the main frames to train our model. Their characteristics are defined as follows:

Episodic Frame

This type of framing approaches the subject by focusing on specific events. The narration of a particular case is presented without context about the causes of the problem and is rather concerned with describing only that event. The treatment of the problem is framed more as an anecdotal event without going deeper into the background of the problem and remains on an individual level (Semetko and Valkenburg, 2000).

Specifically in this study, this framing referred to news stories that specifically emphasised violent acts against women, very often with sensationalist overtones, highlighting stereotypes and prejudices such as, for example, referring to feminicide as a crime of passion (Escribano, 2014). Another way of identifying this framing is when the social problem is reduced to a particular case whose underlying solution falls solely on the victim's direct aggressor without giving more context.

Thematic Frame

The thematic frame, as opposed to the episodic frame, does provide more context and may even contain an analysis of the social problem in question (Semetko and Valkenburg, 2000). The news story, beyond focusing on the individual incident, has a narrative of the social problem in which explanations of possible causes, protocol measures or measures to help eradicate the problem are presented (Escribano, 2014).

Criteria for Case Selection

Gender-based violence permeates, to a greater or lesser extent, in all Mexican states. Because of this, it was established that the states to be compared should have some local digital media with access to a digital newspaper library; both states should reflect high rates of gender-based violence; and they should have opposing legal definitions of femicide, i.e. one of the states should have more legal efforts to combat the crime of femicide - as this is considered the most radical expression of gender-based violence. In addition, it was decisive that one state had an active AVGM and the other did not. This is how the states of Chihuahua and the State of Mexico were chosen, as well as the local digital media *La Opción de Chihuahua* (LOC) and *8 Columnas* (8C).

Data collection technique

The study by Sonia M. Frías (2017) was used as a parameter to decide which keywords to use. The following keywords were chosen for the search of articles on gender violence for the State of Mexico in 2015 and 2019 and for Chihuahua 2019 and also for the download of the new dataset to be categorized:

> woman, girl, feminicide, murdered, dead, raped, disappeared, sexual abuse, beaten, rape, gender violence, violence, dating violence and obstetric violence.

For the present research, the keywords that allowed for a broader enquiry were the terms *mujer*, *niña*, *violencia de género* and *feminicidio*.

Inter-rater reliability

The retrieval of news was saved in a database and processed using the SPSS statistical software. With the help of this statistical package, inter-rater reliability coefficients were obtained. Following the methodology of Krippendorf (2011) the alpha test, defined as follows:

$$\alpha = 1 - \frac{D_o}{D_e}$$

The resulting coefficient indicate the level of agreement between two or more coders using the same tool to assign values to typically unstructured phenomena. Where D_o equals the number of cases where there is disagreement between coders and D_e are the cases of expected disagreement, considering that the correspondence may be due to chance and not to the qualities of the observed phenomenon. Thus, the closer the reliability coefficient α is to 1, the higher the level of inter-coder agreement. Conversely, the closer it is to 0, the less significant it is for the study. For the Frames cathegory (Episodic and Thematic) this the alpha K was 0.89 (Márquez Ampudia, 2023, pg. 54 orig.).

Training Data Description

A total of 972 news items were collected from a systematic search. Of this sample, 73% are news items published by LOC (391 in 2015 and 315 in 2019; 706 in total). The remaining 27% were published by 8 Columns (127 in 2015 and 139 in 2019; 266 in total). *La Opción de Chihuahua* published around 1250 stories per month, close to 7,500 stories in six months. News items covering violence against women and girls represented, on average, 5.2% in 2015 and 3.8% in 2019 of the total number of news items published. 8C, meanwhile, published an average of 365 news items per month, or 2,190 stories in six months. News covering gender-based violence accounted for 5.8% of the total news in 2015 and 6.2% in 2019. The frames of interest for this study were almost equally distributed with a total of 512 Episodic and 460 Thematic news.

New Data Description

To classify news articles on gender-based violence as either *Episódico* or *Temático* frames, a structured natural language processing (NLP) workflow was employed. First, the data, which consisted of 900 labeled news articles, was preprocessed to ensure textual uniformity and accuracy. The article titles were cleaned by converting text to lowercase, removing special characters, diacritical marks (e.g., accents), and redundant whitespace using Regular Expressions. This preprocessing step ensured that the textual data was standardized for further analysis. The cleaned article titles served as input features, while the manually assigned frames acted as target labels.

For the classification task, the Term Frequency-Inverse Document Frequency (TF-IDF) technique was applied with the TfidfVectorizer library from scikitlearn[4] to convert textual data into numerical feature representations (Lane & Dyshel, 2024). Using these vectorized features, a logistic regression model was trained and evaluated. The dataset

[4] https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

was split into training and test sets with an 80-20 split to ensure robust model evaluation. Model performance was assessed using metrics such as accuracy, precision, recall, and F1-score, with a specific focus on the balance between precision and recall for each class (Lewis et al 2022).

Following this process, after training and evaluating the model on the labeled dataset, the model was saved to apply the same approach to classify new, unlabeled news articles. To get the new data there was a scraping process that retrieved a total of 800 news articles, where half of them belonged to 8C news outlet and the other half to LOC. These new articles were then vectorized using the same TF-IDF vectorizer that was fitted on the training data which ensured that the new articles were represented in the same feature space as the training set.

Once the new articles were preprocessed and transformed into the TF-IDF feature space, the trained logistic regression model was used to predict the class (either *Episódico*, 0, or Temático, 1) for each article. The model's predictions were then based on the patterns it learned from the labeled dataset.

Results

Going back to the main question of this research note, it can be stated that the model shows a strong overall performance with an accuracy of 91.2%, correctly classifying most test samples as either *Episódico* or *Temático*. The class-level metrics also highlight the model's effectiveness in distinguishing between these two categories. For the *Episódico* class, the model exhibits high precision (91%) and recall (93%), ensuring both accurate predictions and minimal misclassification. The F1-score of 92% reflects a well-balanced performance. For the *Temático* class, the precision is slightly higher at 92%, but the recall drops to 89%, indicating that while the model is good at identifying *Temático* articles, it misses a few more than it does for *Episódico* articles. The F1-score for *Temátic* is 91%, showing solid overall performance despite this minor discrepancy in recall.

Although the model performs well, it's important to acknowledge that the training data consisted of only 972 news samples, which may still be a limited dataset for certain applications. The macro average of precision, recall, and F1-scores is 91%, reflecting consistent performance across both classes. This metric treats each class equally, regardless of sample size, and shows that the model is equally proficient at classifying both *Episódico* and *Temático* articles. The weighted average, which accounts for class imbalance by weighting metrics based on the number of samples, also equals 91%, as both classes have similar test sample sizes (103 92 respectively). Overall, the model's strong accuracy and balanced class performance demonstrate its effectiveness in categorizing both classes.

Classification Report

	precision	**recall**	**F1-score**	**support**
episodico	0.91	0.93	0.92	103
tematico	0.92	0.89	0.92	92
accuracy			0.91	195
Macro avg	0.91	0.91	0.91	195
Weighted avg	0.91	0.91	0.91	195

Evaluation from the predictive model

After applying the trained model to the new article dataset, the results revealed a notable difference in news categorization, with the Thematic frame being the dominant one across both news outlets (see Figure 1). This could be attributed to a slight underrepresentation of this frame in the training data, which may have led the model to overfit to the more frequently represented category. To address this imbalance in future iterations, incorporating a weighted average into the training process could help better balance the class distribution.

To gain further insights into how different narratives were classified, a Word Cloud was implemented (see Figure 2), providing a clear overview of the main topics identified by the model. As anticipated, and in line with previous research (Márquez Ampudia, 2023), the word "Feminicidio" appeared in both frames. Interestingly, the Episodic frame showed a higher frequency of this term, suggesting that more cases are being accurately recognized as femicides, rather than being mischaracterized as "passional revenge" or other inaccurate labels.

Additionally, the verbs used in both frames align with the respective narratives. For example, terms related to "prosecution" are more closely associated with crime cases, fitting well within the Episodic frame. On the other hand, the term "CODHEM," which refers to the Human Rights Commission of the State of Mexico, suggests a broader, structural approach to gender-based violence, fitting the Thematic frame. This demonstrates that the model is effectively capturing the different ways in which gender-based violence is framed in the media.

Figure 1. Distribution of Frames (Episódico vs Temático) by Newspaper

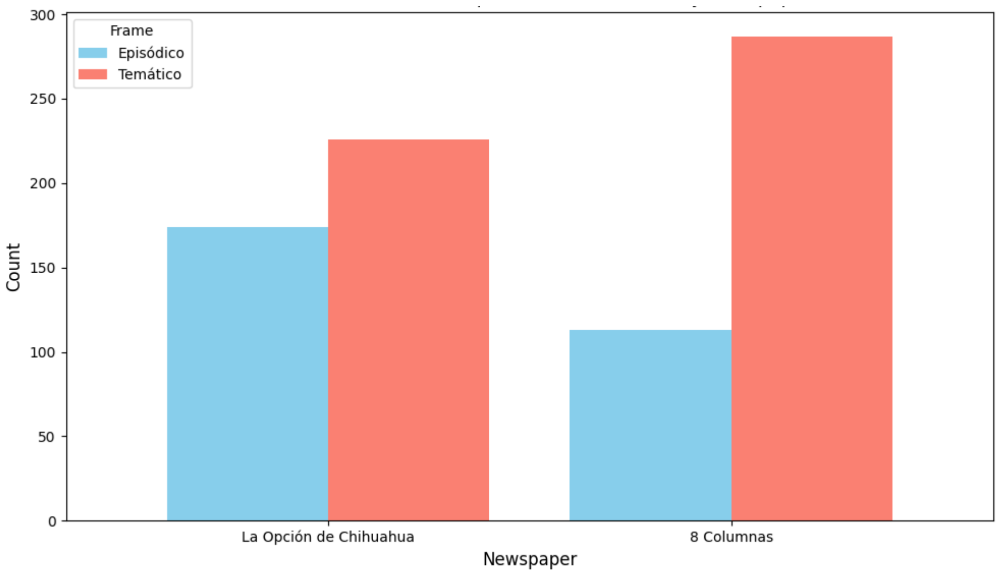
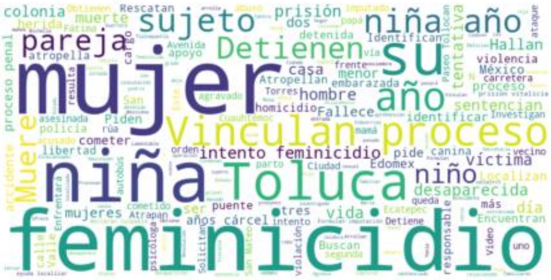
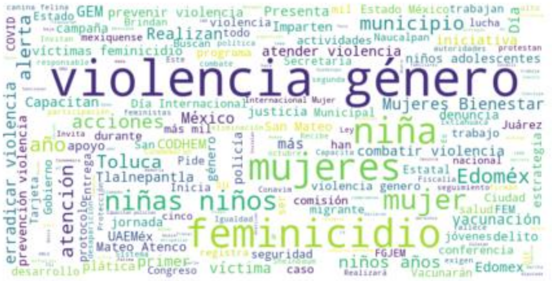


Figure 2. Word Cloud for *Episódico* and *Temático* frames

"Episódico" frame first word cloud



"Temático" frame word cloud



Conclusions

Classifying news coverage, particularly on sensitive topics like gender-based violence, is of critical importance. A model like the one introduced in this research can help to accurately identify and categorize how such issues are presented in the media. This can provide quantitative empirical evidence for media outlest to maybe follow some standarize guidelines that ensure that the portrayal of gender-based violence is approached with the necessary nuance and sensitivity.

Of course to for this amount of detail, further research is needed to explore the mutually inclusive sub-frames within news content, the Political Administrative, Cultural, and Judicial Police sub-frames (Escribano González, 2014, Márquez Ampudia, 2023). These sub-frames are crucial for identifying the various actors and discursive elements present in news narratives, offering a deeper understanding of how different frames shape the interpretation not only of the news, but of the phenomena of gender violence itself. By incorporating these sub-frames into the model, future studies could enhance its ability to analyze the context of this complex news content, leading to more nuanced classifications and insights.

Also, given the resources constrain for this research note, there is more work to be done from retrieving a bigger amount of notes with a larger variaty of keywords, to a more in detail evaluation of the accuracy of the results. So even if with the original training and test data the model showed a strong performance in classifying the *Episódico* and *Temático* frames, with an accuracy of 91.2% and F1-scores of 92% and

91% respectively this should be considered a first phase of a further finetuning of the model. Nonetheles, this note serves as a tool that can function as baseline for future research with a larger and more diverse dataset to improve the model's robustness and its ability to generalize across different news topics.

References

Angélico, R., Dikenstein, V., Fischberg, S. y Maffeo, F. (2014). El feminicidio y la violencia de género en la prensa argentina: un análisis de voces, relatos y actores. Universitas Humanística, *78*, pp. 281-303. doi: 10.11144/Javeriana.UH78.fvgp.

Colombini, M., Mayhew, S.H., Hawkins, B., Bista, M., Joshi, S.K., Schei, B., Watts, C. (2016). Agenda setting and framing of gender-based violence in Nepal: how it became a health issue. *Health Policy and Planning*, *31(4)*. pp. 493–503, https://doi.org/10.1093/heapol/czv091.

Entman, R. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43(4)*. pp. 51 – 58.

Escribano González, M.I. (2014). *Encuadres de la Violencia de Género en la Prensa Escrita y Digital, Nacional y Regional. La Verdad, La Opinión, El Mundo y el País desde la Teoría del Framing" (2005-2010)*. [Tesis doctoral]. Universidad de Murcia. Recuperado en: https://www.tesisenred.net/bitstream/handle/10803/277182/TMIEG.pdf?sequence=1&isAllowed=y.

Goffmann, E. (1974). Frame Analysis. An Essay on the Organization of Experience. Northeastern University Press.

Lane, H., & Dyshel, M. (2024). "Math with words (TF-IDF vectors)", in *Natural Language Processing in Action*. Second Edition. Manning.

Marín, F., Armentia, J.I. y Caminos, J. (2011). El tratamiento informativo de las víctimas de violencia de género en Euskadi: Deia, El Correo, El País y Gara (2002-2009). Comunicación y Sociedad 24 (2), pp. 435-466. URI: https://hdl.handle.net/10171/27354.

Márquez Ampudia, A. (2023). Análisis de la cobertura noticiosa de la violencia de género en México: difusión de hechos violentos y atención al problema en Chihuahua y el

Estado de México (2015 y 2019). Anuario de Investigación de la Comunicación CONEICC, (XXX).

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of communication, 50(2)*, 93-109.

Scheufele, D. A. y Tewksbury, D. (2007). “Framing, Agenda Setting and Priming: The evolution of the three media effects models”. *Journal of Communication*, *57*, pp. 9 – 20.

Starmer, J. (13 March 2023). StatQuest, YouTube. *Word Embedding y Word2Vec, clearIt explained!!!*. https://www.youtube.com/watch?v=viZrOnJclY0.

Lewis, T., von Werra, L. & Wolf, T. (2022). “Text Classification” in *Natural Language Processing with Transformers*. Building Language Applications with Hugging Face. Sebastopol, CA: O’Reilly.

Ulloa Luna, J. L. y Spiller, A. (2014). La cobertura informativa de la violencia de género en medios impresos de Guadalajara, Jalisco, México. El caso de Imelda Virgen. En Libro de Actas del II Congreso Internacional de Comunicación y Género. pp. 641-652. Dykinson.