Detecting Sexism in Tweets: A Sentiment Analysis and Graph Neural Network Approach

Diana P. Madera-Espíndola[1], **Zoe Caballero-Domínguez**[1], **Valeria J. Ramírez-Macías**[1], **Sabur Butt**[1,2], **Hector G. Ceballos**[1,2]

[1]Tecnológico de Monterrey, [2]Institute for the Future of Education

A01025835@tec.mx, A01747247@tec.mx, A01636965@tec.mx

saburb@tec.mx, ceballos@tec.mx

Abstract

In the digital age, social media platforms like Twitter serve as an extensive repository of public discourse, including instances of sexism. It is important to identify such behavior since radicalized ideologies can lead to real-world violent acts. This project aims to develop a deep learning-based tool that leverages a combination of BERT (both English and multilingual versions) and GraphSAGE, a Graph Neural Network (GNN) model, alongside sentiment analysis and natural language processing (NLP) techniques. The tool is designed to analyze tweets for sexism detection and classify them into five categories.

1 Introduction

In today's digital age, social media platforms such as Twitter have become central to public discourse, providing users with a space to express their thoughts and opinions, while also serving as a powerful tool for activism (ElSherief et al., 2017). However, while social media can empower victims to share their experiences, it also enables the spread of harmful ideologies such as sexism and Gender-Based Violence (GBV) (Martínez et al., 2021).

Peter Glick and Susan Fiske introduced a theory in 1996 that explains how power imbalances and mutual dependence between men and women give rise to two interconnected forms of sexist attitudes (Bareket and Fiske, 2023). The first, hostile sexism (HS), is marked by overtly attitudes, including aggression, resentment, objectification, sexual violence, and misogyny. In contrast, benevolent sexism (BS) praises women who conform to traditional roles, offering protection and admiration. However, this attitude is based on the belief that women are inherently weaker, reinforcing harmful stereotypes and gender inequality (Rodríguez-Sánchez et al., 2024).

This dynamic of sexism is not limited to interpersonal interactions but extends to digital platforms, where it takes on new forms and reaches broader audiences. Twitter, with its 280-character limit, often amplifies the problem of hate speech, including gender-based violence and sexism, by encouraging more aggressive and sensational content compared to platforms like Facebook (Founta et al., 2018). Therefore, systems that accurately detect hate speech are crucial for proactive moderation (Davidson et al., 2017).

Sentiment analysis techniques are commonly employed to extract insights about the public sentiment on a wide range of topics, including sexism (Caruccio et al., 2022; Anna Maria Górska and Jemielniak, 2023). When it comes to this classification task, a variety of approaches have been explored, incorporating both machine learning (Sreekumar et al., 2024) and deep learning tools (Castorena et al., 2021; Al-Garadi et al., 2022; Kalra and Zubiaga, 2021). A popular advancement in text representation involves the use of transformers, like BERT, which capture deep, bidirectional contextual information, significantly enhancing the understanding of language complexities (Magnossão et al., 2021; Butt et al., 2021).

However, despite its strengths, these techniques often struggle to capture the complex relationships and structures within texts, such as dependencies between words and tend to underperform when dealing with long-range dependencies. To address these limitations, Graph Neural Networks (GNNs) have been applied in text classification tasks, as they are capable of modeling relationships and dependencies between nodes by propagating information along edges (Khosravi et al., 2024; Utku et al., 2023; Singh and Singh, 2024). Additionally, to further enhance performance, recent approaches have sought to combine BERT embeddings with GNNs (Liu et al., 2025). Our approach builds on this by leveraging BERT's capacity to understand complex language contexts alongside GraphSAGE, a GNN model chosen for its ability to generate

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 48–54

node representations by aggregating features from neighboring nodes (Lu et al., 2024).

Then the contributions of this research in addressing the sexism classification task are summarized as follows:

- The use of representation embeddings combined with GraphSAGE, a GNN model, for detecting and classifying sexism in social media text.
- A competitive tool's accuracy in classifying instances of sexism through a binary classification.
- A comparative study of our proposal against some relevant models in the EXIST 2021 competition.

The rest of the paper is structured with Section 2 covering the dataset and methodology of the proposed model, Section 3 presenting the results, and Section 4 discussing the findings.

2 Method and Data

2.1 Data Description

The dataset used in our study was sourced from the 2021 edition of the sEXism Identification in Social neTworks (EXIST) contest (Rodríguez-Sánchez et al., 2021; Montes et al., 2021), which aims to promote the automatic identification of sexism by providing a benchmark dataset. This dataset includes data from Twitter and Gab.com in both English and Spanish [1]. This distinction highlights the challenge of training a model on one type of structure (tweets) while testing it on a different structure (gabs) to evaluate its adaptability. For this work, we used only the English dataset, which contains 3,436 tweets for training and 2,208 for testing.

The classification task consists of two main subtasks. Task 1 is a binary classification problem, where automated systems must determine whether a message is sexist or non-sexist, as illustrated in Figure 1. The second subtask, shown in Figure 2, involves categorizing a message that has been identified as sexist according to the type of sexism it conveys, such as ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny or non-sexual violence.

[1] https://nlp.uned.es/exist2021/

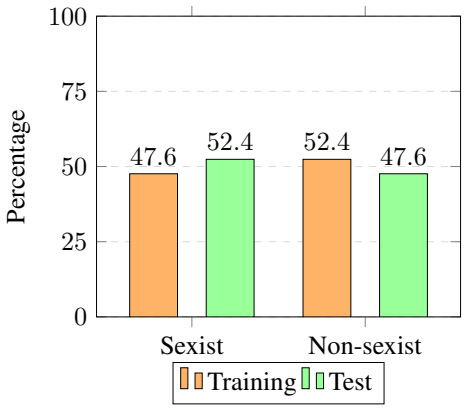

Figure 1: Proportion of Training and Test Datasets for Binary Classification for the EXIST dataset

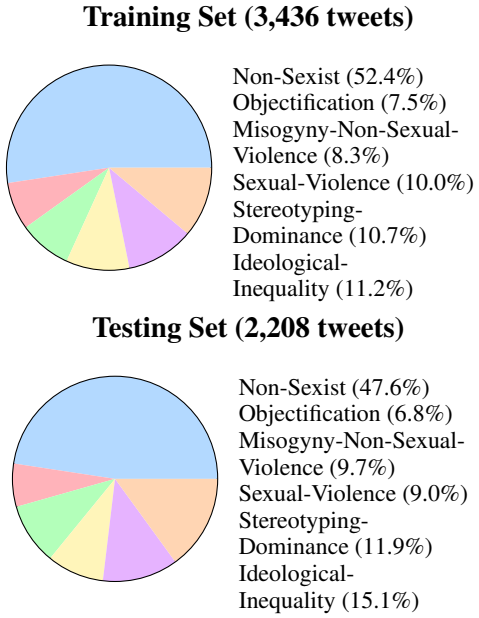

Figure 2: Comparative Class Distribution for English tweets in the EXIST dataset

2.2 Data Processing

In order to enhance the performance of the GraphSAGE. The cleaning process involved:

- Converting text to lowercase
- Removing HTTP links
- Removing Twitter mentions (@username)
- Removing punctuation marks
- Eliminating repeated consecutive letters to at most two consecutive letters
- Removing stop words

2.3 Graph Definition

GraphSAGE is a framework for inductive representation learning on large graphs. It is particularly useful for generating low-dimensional vector representations for nodes, especially in graphs with rich node attribute information (Hamilton et al., 2017). In our case, the graph captures relationships between tweets based on the similarity of their content. First, let us define a graph G as a tuple $G = (V, E)$, where V is the set of nodes (in this case, tweets) of the graph, and E is the set of edges (connections between similar tweets).

For the text numerical representation, we decided to experiment with both the English and Multilingual versions of Bidirectional Encoder Representations from Transformers (BERT and mBERT, respectively). We tested the multilingual version of mBERT to to assess its effectiveness in handling the complexities of multilingual examples, as social media content often contains tokens in multiple languages (Magnossão et al., 2021). Additionally, we included a sentiment polarity attribute because, as noted by (Raees and Fazilat, 2024), it is a key factor in identifying the positive or negative sentiment of a tweet.

Therefore in our graph, each node representing a tweet T_i was associated with four attributes: an embedding vector e_i generated by the pre-trained model, the sentiment polarity score, and two labels. The first label was a binary encoded label, while the second was a multiclass encoded label. Regarding the graph connections, we chose four metrics to appropriately weigh the edges, with the goal of forming a composite weight. Two of these metrics, include cosine similarity between the tweet embeddings and cosine similarity between TF-IDF vector representations (Nakajima and Sasaki, 2023). We decided to incorporate the vector representation TF-IDF to complement the embeddings, as it is a statistical measure used to evaluate the importance of a word in a document relative to a corpus (Khosravi et al., 2024).

For the remaining two metrics that contribute to the composite weight, we chose semantic similarity and sentiment agreement. For semantic similarity, we used the NLP model en-core-web-md from $SpaCy$, which computes the similarity between the embeddings of the tweets. For sentiment agreement, we used the sentiment polarity score from $TextBlob$ to calculate the sentiment of each tweet. The sentiment agreement is then determined by calculating

$$1 - abs(sent1, sent2) \tag{1}$$

where $sent1$ and $sent2$ represent the sentiment polarity scores of the two tweets being compared.

To identify the optimal weights for each metric, we conducted two experiments on the training set. One experiment used nodes generated by BERT, while the other used nodes generated by mBERT. For each pair of nodes, we calculated four key metrics: cosine similarity between embeddings, cosine similarity between TF-IDF vectors, semantic similarity, and sentiment agreement. To optimize memory usage, the training dataset was divided into smaller chunks during computation.

After calculating the four metrics, they were normalized to ensure compatibility with the Louvain algorithm. This algorithm partitions a network into communities by first assigning each node to its own community, then iteratively merging nodes or communities to maximize modularity. By optimizing modularity, the algorithm identifies clusters where nodes are more strongly connected to each other than to those outside the cluster (Kim and Sayama, 2019).

We tested 15 random weight combinations, each prioritizing a specific metric, to assess its impact on community formation. This approach enabled us to evaluate the importance of each metric in creating meaningful communities. Finally, we analyzed the results to determine the weight combination that produced the most cohesive community structure, using modularity as the evaluation criterion. Based on the experiment with the highest modularity score, we assigned the following weights:

$$\begin{aligned}\text{composite weight} = {} & \text{cosine similarity TF-IDF} \times 0.1 \\ & + \text{semantic similarity} \times 0.8 \\ & + \text{sentiment agreement} \times 0.1\end{aligned} \tag{2}$$

Additionally, to reduce noise and avoid computational problems due to a very dense graph, we established a threshold of 0.7, ensuring that only edges with a similarity score above this threshold are created. Furthermore, we limited the number of connections per node to a maximum of 5. Finally, we construct the graph by creating an Adjacency matrix A, where each entry A_{i_j} corresponds to the edge between tweets T_i and T_j.

2.4 GraphSAGE

Unlike previous approaches that require all nodes to be available during the training of embeddings,

GraphSAGE leverages node feature information to create effective representations even for unseen nodes. This inductive property allows the algorithm to generalize beyond the trained data (Hamilton et al., 2017). As demonstrated by (Lu et al., 2024), integrating BERT into the GraphSAGE framework significantly improves generalization ability and classification accuracy compared to traditional graph-based and BERT-based models. While their study focused on classification within a citation network, they also tested their model on sentiment analysis for movie reviews, which motivated us to explore this GNN model for our task.

For Task 1, we used two layers with ReLU activation and a Sigmoid function for the output layer. The Adam optimizer was employed, with Binary Cross-Entropy as the loss function. For Task 2, we also used two layers with ReLU activation, but applied the argmax function for the output. The Adam optimizer was retained, and the loss function was changed to Cross-Entropy for multiclass classification. To address class imbalance, we assigned higher weights to less frequent classes, ensuring the model focused more on these during training.

2.5 Hyperparameter Optimization

We used Optuna (Akiba et al., 2019), a framework for hyperparameter optimization, targeting validation accuracy, which allowed the framework to iteratively test various configurations and select the best. The key hyperparameters optimized were hidden channels, dropout rate, learning rate, weight decay, and epochs. These were selected because hidden channels enhance feature learning, dropout rate helps reduce overfitting, learning rate and weight decay balance convergence and regularization, and epochs control the training depth and efficiency.

We tested two different graphs for both tasks: one with BERT embeddings, and another one with mBERT embeddings. We ran 100 Optuna trials for each one of the four models. Table 1 show the obtained best configurations for the hyperparameters. An important note is that we used a transductive training approach, where the training, validation, and test sets are part of the same graph but segmented through attribute-based masking. This setting enables us to leverage all available node information within the graph structure (Li et al., 2021).

Task	Hyperparameter	BERT graph	mBERT graph
Task 1	**Hidden channels**	62	62
Task 1	**Dropout rate**	0.178	0.1837
Task 1	**Learning rate**	0.0032	0.00078
Task 1	**Weight decay**	0.0049	0.0056
Task 1	**Epochs**	76	102
Task 2	**Hidden channels**	128	128
Task 2	**Dropout rate**	0.4170	0.5009
Task 2	**Learning rate**	9.1012	9.9856
Task 2	**Weight decay**	0.0061	0.0052
Task 2	**Epochs**	320	300

Table 1: Best hyperparameter configuration obtained by Optuna for both graphs in Task 1 (binary classification) and Task 2 (multiclass classification)

Task	Model	Accuracy	F1
Task 1	BERT	0.7020	0.7303
	mBERT	0.6359	0.6271
Task 2	BERT	0.5308	0.3783
	mBERT	0.5231	0.2981

Table 2: Performance Comparison of Both Models

3 Results

Results of both tasks are presented in Table 2 showing a comparison of the main metrics obtained on Task 1 (binary) between the first proposed model, which uses embeddings generated with BERT, and the second model, which uses embeddings generated with mBERT. It is important to note that the primary metric we are using to measure the success of our model is the F1-score.

Actual	**Predicted**	
	Sexist	**Not Sexist**
Sexist	True Positive 840	False Negative 318
Not Sexist	False Positive 340	True Negative 710

Table 3: Confusion Matrix for BERT Embeddings Model on Task 1

The BERT graph exhibited strong performance in distinguishing sexist tweets from non-sexist ones, achieving an accuracy of 0.702 and an F1 score of 0.731. In contrast, mBERT produced lower

results for both metrics, highlighting the superiority of BERT over mBERT on this task. Table 3, presents the confusion matrix for the BERT model in this binary classification, showing similar error rates for false positives and false negatives. Although, the model slightly favors non-sexist classification, with 318 false negatives compared to 340 false positives, indicating a relatively balanced performance.

Act. \ Pred.	NS	II	SD	OBJ	SV	MNSV
NS	413	172	150	104	118	93
II	56	203	38	13	14	9
SD	45	62	87	28	25	15
OBJ	15	5	31	70	20	9
SV	16	24	22	25	82	29
MNSV	29	36	34	27	24	65

Table 4: Confusion Matrix for BERT Embeddings Model on Task 2. NS: Non Sexist, II: Ideological Inequality, SD: Stereotyping Dominance, OBJ: Objectification, SV: Sexual Violence, MNSV: Misogyny Non Sexual Violence

For Task 2, both embeddings showed lower performance overall; however, BERT continued to demonstrate its advantage over mBERT with an accuracy of 0.5308 and an F1 score of 0.3783. Table 4, displays the confusion matrix for BERT in this multiclass classification task, revealing the highest confusion between the Non-Sexist (NS) and Ideological Inequality (II) classes, with 172 instances of II misclassified as NS. There was also significant confusion between NS and Stereotyping Dominance (SD), with 150 misclassifications. Overall, the model shows a bias toward classifying instances as NS but performs best at identifying the Ideological Inequality (II) class, with a precision of 61.0%. It struggles the most with the MNSV (30.2% precision) and SD (33.2% precision) classes.

4 Discussion and Related Work

The application of GNNs, such as GraphSAGE, to text classification, remains relatively unexplored but holds considerable promise. Our model demonstrated competitive performance on Task 1. However, it requires improvements for Task 2.

The first-place team in the EXIST contest (Magnossão et al., 2021) created a second version of both BERT and mBERT by translating some instances from Spanish to English to enhance the training data. They also implemented ensemble strategies, combining predictions from individual models, which consistently outperformed the single mBERT and BERT models. Therefore, integrating some data strategies and an ensemble of GraphSAGE networks could be a worthy experiment for future research. Nonetheless, this entry was not the only using data augmentation strategies. Butt et al. (Butt et al., 2021) used a 'Back Translation' strategy, where they input the text in the source language, translate it to another second language, and finally back to the source language. Furthermore, data augmentation strategies can also be utilized to mitigate the class imbalance problem of Task 2.

Among the models reviewed by the contestants, MB-Courage (Wilkens and Ognibene, 2021) was the model most closely aligned with our proposed approach, as it also utilizes GNN for identifying sexism. However, while MB-Courage employs Graph Convolutional Neural Networks (GCN), we use GraphSAGE, a distinct variation of GNN. In terms of performance, our model outperformed MB-Courage's best proposal on F1-score for Task 1. Regarding Task 2, our best proposal proved to be the least effective among the compared models showed in Table 5. We attribute the low performance in this second task to class imbalance and the model's difficulty in understanding the context of statements. This explains why it can generalize for two classes but struggles to adapt to multiclass classification.This would also explain why proposal that performed data augmentation performed well. By adding more examples of each class, the class imbalance could be lessen and, in turn, the model may enough data to distinguish the different classes.

As a final point, exploring alternative text similarity metrics such as emotion detection, Latent Dirichlet Allocation (LDA) topic modeling, or ConceptNet similarity, could provide valuable insights for defining the graph structure, leading to improved performance in second tasks. Moreover, improved text preprocessing and experimenting with different embedding models could help preserve higher-quality information.

Conclusion and future work

Our best model archived an F1 score of 0.7331 on Task 1, which demonstrates competitive performance, as this would have placed us 29th out of

Task	Model	Accuracy	F1 Score
Task 1	**mBERT & GraphSAGE**	0.636	0.627
Task 1	**BERT & GraphSAGE**	0.702	0.730
Task 1	**Ensemble Model**	0.789	0.780
Task 1	**GCN**	0.715	0.715
Task 1	**BERT & Data Augmentation**	0.728	0.727
Task 2	**BERT & GraphSAGE**	0.531	0.378
Task 2	**mBERT & GraphSAGE**	0.523	0.298
Task 2	**Ensemble Model**	0.658	0.579
Task 2	**GCN**	0.595	0.459
Task 2	**BERT & Data Augmentation**	0.553	0.491

Table 5: Accuracy and F1 scores of our models for Task 1 (binary classification) and Task 2 (multiclass classification), compared to those reported by Magnossão de Paula et al., Wilkens & Ognibene, and Butt et al.

72 participants in the competition, and also outperforms the only Graph Neural Network proposal used in the competition. This performance shows the potential of using Graph Neural Networks for sexism in text settings.

However, further enhancements can be made to improve upon these results, especially in Task 2, where our model only managed an F1 score of 0.378 (56th place out of 72). Exploring data augmentation techniques and incorporating an ensemble of GraphSAGE networks could be valuable, particularly for tasks like Task 2, where class imbalance was a significant factor. Additionally, experimenting with different text similarity metrics and enhancing data pre-processing approaches could lead to better performance.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, 15:100217.

Karolina Kulicka Anna Maria Górska and Dariusz Jemielniak. 2023. Men not going their own way: a thick big data analysis of #mgtow and #feminism tweets. *Feminist Media Studies*, 23(8):3774–3792.

Orly Bareket and Susan T Fiske. 2023. A systematic review of the ambivalent sexism literature: Hostile sexism protects men's power; benevolent sexism guards traditional gender roles. *Psychological bulletin*.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander F Gelbukh. 2021. Sexism identification using bert and data augmentation-exist2021. In *IberLEF@ SEPLN*, pages 381–389.

Loredana Caruccio, Stefano Cirillo, Vincenzo Deufemia, Giuseppe Polese, and Roberto Stanzione. 2022. Data analytics on twitter for evaluating women inclusion and safety in modern society. In *itaDATA*.

Carlos M. Castorena, Itzel M. Abundez, Roberto Alejo, Everardo E. Granda-Gutiérrez, Eréndira Rendón, and Octavio Villegas. 2021. Deep neural network for gender-based violence detection on twitter messages. *Mathematics*, 9(8).

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Mai ElSherief, Elizabeth Belding, and Dana Nguyen. 2017. #NotOkay: Understanding gender-based violence in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1).

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Preprint*, arXiv:1802.00393.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Amikul Kalra and Arkaitz Zubiaga. 2021. Sexism identification in tweets and gabs using deep neural networks. *arXiv preprint*.

Asal Khosravi, Zahed Rahmati, and Ali Vefghi. 2024. Relational graph convolutional networks for sentiment analysis. *arXiv preprint*.

Minjun Kim and Hiroki Sayama. 2019. The power of communities: A text classification model with automated labeling process using network community detection.

Chen Li, Xutan Peng, Hao Peng, Jianxin Li, and Lihong Wang. 2021. Textgtl: Graph-based transductive learning for semi-supervised text classification via structure-sensitive interpolation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, page 2680–2686, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Qi Liu, Kejing Xiao, and Zhaopeng Qian. 2025. A hybrid re-fusion model for text classification. *Scientific Reports*, 15(1):9333.

Junwen Lu, Lingrui Zheng, and Moudong Zhang. 2024. Application of bert-graphsage model in text and paper classification tasks. In *Advanced Data Mining and Applications: 20th International Conference, ADMA 2024, Sydney, NSW, Australia, December 3–5, 2024, Proceedings, Part V*, page 315–327, Berlin, Heidelberg. Springer-Verlag.

Angel Felipe Magnossão, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *CoRR*, arXiv:2111.04551.

Fátima Martínez, Carolina Pacheco, and Marco Galicia. 2021. The #metoo movement in twitter: Fighting gender-based violence. In *Information Technology and Systems*, pages 36–44, Cham. Springer International Publishing.

M. Montes, P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M.A. Álvarez Carmona, E. Álvarez Mellado, J. Carrillo-de Albornoz, L. Chiruzzo, L. Freitas, H. Gómez Adorno, Y. Gutiérrez, S.M. Jiménez-Zafra, S. Lima, F.M. Plaza-del Arco, and M. Taulé. 2021. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*.

Hiromu Nakajima and Minoru Sasaki. 2023. Text classification based on the heterogeneous graph considering the relationships between documents. *Big Data and Cognitive Computing*, 7(4).

Muhammad Raees and Samina Fazilat. 2024. Lexicon-based sentiment analysis on text polarities with evaluation of classification models. *Preprint*, arXiv:2409.12840.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. Detecting sexism in social media: an empirical analysis of linguistic patterns and strategies. *Applied Intelligence*, 54(21):10995–11019.

Francisco Rodríguez-Sánchez, Laura Plaza Jorge Carrillo-de Albornoz, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Loitongbam Gyanendro Singh and Sanasam Ranbir Singh. 2024. Sentiment analysis of tweets using text and graph multi-views learning. *Knowledge and Information Systems*, 66(5):2965–2985.

Murari Sreekumar, Shreyas Karthik, Durairaj Thenmozhi, Shriram Gopalakrishnan, and Krithika Swaminathan. 2024. Sexism identification in tweets using machine learning approaches. In *Conference and Labs of the Evaluation Forum*.

Anıl Utku, Can Umit, and Aslan Serpil. 2023. Detection of hateful twitter users with graph convolutional network model. *Earth Science Informatics*, 16(1):329–343.

Rodrigo Souza Wilkens and Dimitri Ognibene. 2021. Mb-courage @ exist: Gcn classification for sexism identification in social networks. In *IberLEF@SEPLN*.