IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 December 2022, accepted 21 January 2023, date of publication 7 February 2023, date of current version 13 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242965

RESEARCH ARTICLE

Enhancing the Detection of Misogynistic Content in Social Media by Transferring Knowledge From Song Phrases

RICARDO CALDERÓN-SUAREZ[1], ROSA M. ORTEGA-MENDOZA[1], MANUEL MONTES-Y-GÓMEZ[2], CARINA TOXQUI-QUITL[1], AND MARCO A. MÁRQUEZ-VERA[3]

[1]División de Investigación y Posgrado, Universidad Politécnica de Tulancingo (UPT), Tulancingo, Hidalgo 43629, Mexico
[2]Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla 72840, Mexico
[3]Departamento de Mecatrónica, Universidad Politécnica de Pachuca (UPP), Zempoala, Hidalgo 43830, Mexico

Corresponding author: Rosa M. Ortega-Mendoza (rosa.ortega@upt.edu.mx)

This work was supported in part by the Universidad Politécnica de Tulancingo (UPT); and in part by the Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico, under Grant Scholarship CVU-714747.

ABSTRACT Misogyny is a serious social problem that affects the mental and physical health of women and can even lead to femicide. This problem is visible and prevalent in different communication channels, such as music and social networks, encouraging and reinforcing this harmful behavior. Given this situation, the automatic detection of misogynistic content on social networks is a task of increasing interest. In this regard, most current computational approaches employ a supervised machine learning strategy. The main challenge is to capture the diversity and complexity of offensive language directed at women. Accordingly, the size and quality of training data play a fundamental role in the results of the methods. In this paper, we propose a novel data augmentation approach that takes advantage of song lyrics to increase the generalization capability of methods and improve their performance. Hence, we present a methodology for automatically compiling a corpus of song phrases that show abusive and explicit words against women. The proposed approach was evaluated using English and Spanish benchmark datasets, obtaining results that outperform conventional transfer learning techniques and achieve high competitiveness compared with state-of-the-art methods.

INDEX TERMS Data augmentation, misogyny detection, transfer learning, social media, song lyrics.

I. INTRODUCTION

Misogyny is a serious social problem manifested through cultural beliefs that consider women as inferior beings [1], [2], violating the principle of gender equality, a fundamental and inalienable human right [3]. In general, it has been associated with attitudes biased against women, such as male privilege, gender discrimination, sexual objectification, verbal aggressiveness, and physical violence [2]. These manifestations affect women's self-esteem and can progressively damage their mental and physical health, leading in severe cases even to femicide [4]. Unfortunately, misogyny is visible and prevalent in social media, making them unsafe and unequal spaces for women [5], [6]. Hence, the automatic detection of misogynistic content on these platforms is of increasing concern.

In general, the Automatic Misogyny Identification (AMI) task has been framed as a text classification problem aimed at detecting traces of misogynistic content in social media [7], [8], [9], with the idea that language is related to social phenomena, including this behavior [10], [11]. The complexity of the task is due to the challenge of identifying this type of language both in its explicit and implicit forms [12]. The former involves the recognition of informal vocabulary (e.g., slang words) and the variety of meanings of some keywords depending on their context (e.g., insults). The latter is more complicated because of the use of complex linguistic structures such as humorous (e.g., jokes) and sarcastic or ironic expressions [13], [14]. To address these problems, we propose enriching the models' generalization ability by

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

transferring knowledge from song lyrics. Although the use of words may differ in social media and music, we argue that misogyny in both is closely related because it derives from sociocultural attitudes that might include biases against women. For example, sexual content and derogatory words often have negative connotations for women in both domains.

In particular, we used song lyrics as a source domain to diversify the instances in the training of the target domain defined by the AMI task. To the best of our knowledge, the potential of lyrical content as a unique data source (out-domain) for transferring knowledge into the task has not yet been explored. It is worth mentioning that song lyrics have been previously used as part of a multi-source dataset to extract features for the task at hand [15]. However, that work used song lyrics, mostly synthetic, together with data from other sources, such as documents and proverbs, without discussing the contribution of each one. Moreover, relevant sentences to create the classification model were manually selected using human annotators. In contrast to this previous effort, we propose a fully automatic approach of data augmentation (DA) to transfer knowledge from song phrases to detect misogyny in social media.

This research poses three key research questions. First, do song lyrics contain information that can be exploited to enhance the detection of misogyny in social networks? Second, what instances are best for training a classifier for the AMI task, the entire songs or just a selected subset of their phrases? Third, does the proposed DA approach achieve better results in the task than other basic transfer learning techniques? Investigating these questions, this paper provides the following main contributions: i) A methodology to automatically compile corpora of song phrases containing misogynistic content in English and Spanish, which, in the future, can be used to build resources to enrich the detection of other types of hate speech, also reflected in the lyrics of some songs; ii) A data augmentation approach that leverages song phrases to increase the effectiveness and robustness of methods for misogyny detection in social media.

The rest of the paper is organized as follows. Section II describes previous work on misogyny detection and research on music mining. Section III introduces a methodology to build a corpus of song phrases containing misogynistic content. Section IV presents the proposed DA approach to leverage knowledge contained in song lyrics toward the AMI task. Section V defines the experimental settings that support the experiments. Section VI reports and discusses the results. Finally, Section VII exposes the conclusions and future work.

II. RELATED WORK

A. MISOGYNY DETECTION IN SOCIAL MEDIA

Over the years, misogyny has evolved as an ideology, which is held according to cultural and social contexts [2]. Nowadays, this behavior is exposed in online environments such as micro-blogs due to the freedom to express opinions and sentiments [16]. Unfortunately, expressions of hostility and hate speech against women on social media can be morally harmful and even incite physical violence [2]. For example, a social study found a relation between the number of rapes and misogynistic tweets per state in the USA [17].

In particular, it has been established a strong relationship between misogyny and language [10]. This link has motivated the development of several automatic approaches for the detection of misogynistic comments and posts. In this context, [18] presented a preliminary study on the use of language in misogynistic tweets. Based on this study, in [8] its authors reported the first attempt to automatically detect misogynistic language in social media. To date, diverse methods have tackled this task. A recent review on this topic distinguishes two principal approaches [14]: models that use classical machine learning and those based on Neural Networks. The latest trends use transfer learning [19], especially through pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) [20].

1) SHARED TASKS ON MISOGYNY DETECTION

In 2018, the IberEval forum held a shared task to evaluate automatic methods to detect and categorize misogyny in English, and Spanish tweets [7]. The participating systems used n-grams and embedding-based representations with several linguistic characteristics (e.g., stylistic, structural, lexical, and affective features), specific terms related to swearing words, sexist slurs, and woman-related words. Support Vector Machines (SVM) and ensembles were the classifiers most commonly used. In the same year, a similar forum was proposed for detecting misogyny in both Italian and English tweets [21]. A year later, in 2019, another competition on a related topic was released. It was focused on hate speech against immigrants and women under a multilingual framework [22].[1] In 2020, the AMI task was relaunched with benchmark datasets in Italian [23].

The recent competitions about misogyny detection have released labeled datasets of texts written in distinct languages, such as Spanish and English. The availability of these corpora has encouraged the development of new methods to study, model, and detect misogynistic content.

2) TRANSFER LEARNING IN THE TASK

Recently, transfer learning mechanisms have been explored in the AMI task, especially to face the lack of labeled data. In this regard, basic domain adaptation approaches have allowed leveraging knowledge across domains [24], [25]. For example, datasets about different abusive phenomena (sexism, hate speech, and offensive language) have been used to detect misogyny through cross-domain classification methods [26]. Also, general and specialized pre-trained word embeddings (e.g., those trained using Wikipedia) [27] have been leveraged as a transfer learning approach to

[1] It should be noted that hate speech is directed towards an individual or a group of people based on their characteristics such as gender, religion, race, skin color, among others [22]. In this case, the targeted categories were women and immigrants.

feed neural network models in the task [28], [29]. In this work, we adapted cross-domain and embedding-based methods to transfer knowledge from lyrical content into the AMI task.

Transfer learning is also prevalent through the use of large-scale pretrained language models and fine-tuning them on this downstream task. For example, authors in [19] pretrained an LSTM-based language model on multiple datasets and fine-tuned it to detect misogyny. Recently, BERT-based language models [20] have been applied in the task by employing versions such as DistilBERT, BETO, and DeBERTa [30], [31], [32]. We also exploit these models to assess the proposed DA strategy.

Commonly, DA is oriented to generate synthetic sentences from a training source dataset [33], making them dependent on the quality of the data generated. Instead, we used sentences from lyrics that capture the idiosyncrasies of society. To our knowledge, the use of song lyrics as a single source for data augmenting has not yet been explored. In a previous work [15], distinct datasets, including songs (mostly synthetic), were used to collect misogynistic sentences, which were then ''manually'' inspected and used to build a misogyny classifier. In contrast, this paper explores the contribution of song phrases as a unique data source in a DA approach, and for this purpose we propose an automatic mechanism to select the phrases that appear to be the most relevant to the task.

B. SONG LYRICS AS A KNOWLEDGE SOURCE

Music has been considered a language that conveys meanings because it expresses ideas, feelings, and emotions [34], [35], becoming a powerful communication channel with broad diffusion [36]. Some studies indicate that music is linked to the context where it is produced, establishing an inter-relationship between music, society, and culture [37]. Therefore, the music encompasses a valuable domain that has motivated diverse research.

1) MUSIC MINING AND KNOWLEDGE TRANSFER

Since song lyrics convey messages and emotions [38], they represent a valuable source to study the use of language. Several research works have leveraged this knowledge to perform tasks in an intrinsic way (i.e., in the same domain). For example, some studies have used song lyrics to train word embeddings, which are evaluated in tasks related to music mining, such as emotion detection [39], genre classification, explicit content identification [40], era detection [41], and gender biases [42].

Knowledge from song lyrics has been less exploited for extrinsic downstream tasks. In this scenario, the authors of [43] trained word embeddings from song lyrics combined with social media posts and code-mixed texts (i.e., those with more than one language in its discourse). These word representations were evaluated on two out-of-domain tasks: word analogies and language detection. Although these studies have demonstrated the potential of song lyrics to enrich methods in downstream tasks, to the best of our knowledge, the use of lyrics as a single knowledge source has not yet been studied in the detection of misogyny language.

2) MUSIC AND MISOGYNY

Music has been exploited to study stereotypical roles [44] and society's biases, particularly those against women [42]. Recently, research has explored the relationship between verbal misogyny and song lyrics concluding that several of them portray women negatively [45]. In general, it has been established that music can cause or confirm misogyny [46]. Although manifestations of misogyny are present in most music genres, it is more commonly visible in hip-hop [47], [48], [49], rap [50], [51], [52], metal [46], [53], and country music [54], [55]. Misogynistic songs often contain depictions of women that express rooted phenomena in society, such as verbal sexual objectification of females [50], references to genitalia and other body parts, female inferiority, and physical male violence against women, including rape and murder [56], [57]. Accordingly, we consider that certain song lyrics could be a valuable data source to enrich the training of classifiers for different abusive phenomena, especially the AMI task.

III. BUILDING A CORPUS OF SONG PHRASES WITH MISOGYNISTIC CONTENT

Inspired by findings indicating that several song lyrics expose diverse harmful expressions to women, we propose a methodology to automatically build a dataset of song phrases suitable to support the AMI task. We distinguish two stages in our methodology: gathering misogynistic song lyrics and extracting short explicit abusive phrases from their content. Figure 1 illustrates these two stages, and the following subsections describe the processes involved in them.

A. GATHERING MISOGYNISTIC SONG LYRICS

In this first stage, misogynistic and non-misogynistic songs are collected from the web. To favor linguistic diversity, we gathered song lyrics of different authors and genres using a variety of web sites.[2] Then, to automatically determine the category of a song, we propose to analyze the presence of seed words that come from previous related studies. Two types of seed words were particularly used:

- Misogynistic words: these are words associated with verbal abuse against women. In this research, we employed terms from the misogynistic lexicons described in [16][3] for English, and in [58][4] for Spanish.[5]
- Words related to women: To ensure that song lyrics focus on women, we considered some words commonly

[2]For example, https://www.lyrics.com/ and https://www.letras.com/

[3]https://github.com/miriamfs/WebSci2019; Since it includes general hate words and phrases, we selected only unigrams referring to women

[4]https://github.com/fmplaza/hate-speech-spanish-lexicons

[5]The plural form of these words was also considered.

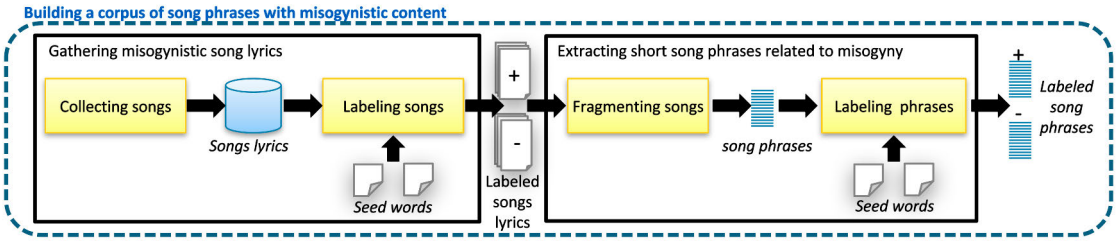


FIGURE 1. Process for the automatic construction of a corpus of song phrases with misogynistic content.

associated with them, for example, *girl*, *girlfriend*, and *wife*.[6]

Specifically, we determined that a song has misogynistic content when it includes words of both types of seeds, and at least two related to misogyny. Conversely, a song is labeled as non-misogynistic if it does not have any misogynistic words, even if it contains terms related to women. After the automatic labeling, we preprocessed them by removing expressions that define the structure of the songs rather than being part of their lyrics (e.g., the phrases *chorus* and *repeat n times*). Besides, the repeated verses were kept only once.

B. EXTRACTING SHORT SONG PHRASES RELATED TO MISOGYNY

Aware of the diversity of content in song lyrics, we decided only to explore the relevance of song phrases with explicit misogyny. Considering that the target application is the detection of misogynistic tweets, we divided the song lyrics into segments with a maximum length of 280 characters, similar to the length of the posts on this platform. Each of these phrases (segments) was automatically labeled according to the following criteria:

- Misogynistic (positive class). They come from songs labeled with this category and include two misogynous keywords[7] and one related to women.
- Non-misogynistic (negative class). They are random phrases that come from songs labeled with this category.

This automatic methodology, in addition to reducing the cost and subjectivity of the labeling, produces good examples of texts with misogyny expressions. This is mainly due to two reasons. First, the song lyrics reflect cultural attitudes that portray society. This is important for both positive and negative categories. Second, they contain high linguistic diversity and therefore combine implicit and explicit misogynistic language.

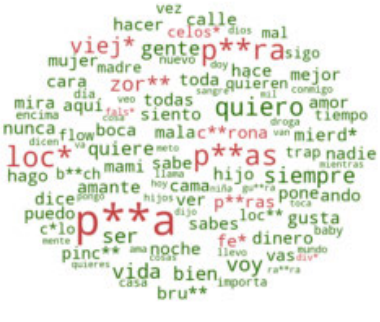
C. RESULTING COLLECTIONS

Table 1 shows some data about the songs collected (Lyrics) and the phrases (SgPh) extracted from them for our English and Spanish datasets, respectively. It is important to note that the two collections were intentionally balanced with respect to the number of instances in each category.

TABLE 1. General statistics of our English and Spanish datasets of song lyrics.

Language	Dataset	Misogynistic	Non-misogynistic	Total
Spanish	Lyrics	4228	4228	8456
	SgPh	1411	1411	2822
English	Lyrics	12117	12117	24234
	SgPh	2120	2120	4240

To analyze the context captured by the extracted song phrases, we explored their vocabulary. Figure 2 shows the top-frequent words in each lyrics collection through word clouds, where the font size is related to the words' frequencies. Seed words are indicated in red and the rest



(a) Spanish

(b) English

FIGURE 2. Top-100 frequent terms in the collected datasets. Seed words are indicated in red. In some cases, we mask words using a "*" to soften their offensiveness.

[6]We obtained words related to women through two sites: https://relatedwords.org and http://www.ideasafines.com.ar/do-buscar.php, in English and Spanish, respectively

[7]We selected two keywords to ensure harmful content against women.

in green. From these word clouds, it is possible to observe that several words are common in both music and social media, especially misogynistic seed words. It is also interesting to note the presence of other terms related to aggressive contexts, such as swearing and derogatory words. Some words referring to parts of the body are also observed, which are closely related to the sexual objectification of women (e.g., **ss* in English and *c*lo* in Spanish). This analysis shows evidence that these small song segments surrounding the seed words capture different manifestations of misogyny.

IV. CROSS-DOMAIN DATA AUGMENTATION USING SONG PHRASES

The AMI task in social media is commonly tackled as a supervised text classification problem. From this perspective, the quality of classifiers is usually related to their generalization ability, which highly depends on the amount of training data. Unfortunately, labeled examples are typically scarce for this domain. As a solution to this problem, and motivated by the prevalence of misogyny in several song lyrics, we propose a data augmentation approach that aims to exploit the knowledge and patterns from these lyrics. Figure 3 shows a general overview of this approach. Its goal is to use high-quality phrases of the song lyrics to augment the training sets related to the task. Hence, its key idea is to increase the learning ability of models by diversifying the training instances by considering examples of socio-cultural expressions from music.

A. TRANSFER LEARNING FROM SONG PHRASES

The proposed approach can be considered a transfer of learning technique due to the intention to leverage existing knowledge in song lyrics for its use in an out-domain task. Based on concepts and notations in [25], [59], and [60], let D_S and D_T be the source and the target domain data, respectively. Also, let T_S and T_T represent the source and target learning task. Transfer learning is aimed at using knowledge in D_S and T_S to help improve the learning process in the target domain, where $D_S \neq D_T$ or $T_S \neq T_T$. In this research, two domains are involved: song lyrics and tweets, considered as D_S and D_T, respectively. Whereas, $T_S = T_T$ since we focus on the identification of misogyny for both domains. Hence, the proposed approach augments training data in D_T following a cross-domain framework. The aim is to aggregate only labeled song phrases that enhance the classifiers' performance. To achieve this, the approach uses the mechanisms described in the following section.

B. FILTERING NOISY SONG PHRASES

DA generally refers to generate additional artificial instances (i.e., synthetic) to increase models' generalization capabilities [33]. However, some noisy instances can be introduced by this process, thereby affecting the effectiveness of the models. Therefore, similar to other previous works [33], we consider necessary to carry out a quality assessment of the augmented data in order to reduce the insertion of low-quality instances. In particular, we propose a similarity-based filter that allows selecting a subset of "reliable" song phrases, which might be relevant to the target task.

The proposed filter is represented in Fig. 4. It aims to retain only the source instances (i.e., song phrases) nearest to the centroids of each class in the target domain (i.e., tweets). In the figure, positive and negative tweets and song phrases are exemplified by small blue and red circles and squares, respectively. To determine the proximity between song phrases and tweets' centroids, we used the *cosine* similarity; then, we retained a percentage (θ) of the most similar song phrases. As an alternative approach, we also employed the Roccio classifier.[8]

Finally, the filtered song phrases are added to the training set of the target domain, and an enhanced classification model is built. It is worth noting that this automatic filter minimizes high divergences between domains and also avoids human assistance to select the best examples in the DA process.

V. EXPERIMENTAL SETTINGS

A. DATASETS

The proposed approach was evaluated on three benchmark collections from the AMI task. In particular, we used the English and Spanish datasets from the IberEval 2018 as well as the English collection from Evalita 2018, hereafter denoted as *Iber-Sp*, *Iber-En*, and *Eva-En*, respectively. Table 2 presents the general distribution of these datasets. The collections contain tweets labeled as misogynistic and non-misogynistic, referenced as the positive and negative classes, respectively. Tweets labeled as positive are very diverse, some contain explicit expressions with abusive language against women, while others are equally offensive but more subtle, using sarcasm or humor, as in the following example: "*How many men are required to clean a house? Zero, that's a women's task*".[9]

TABLE 2. Data distribution of the collections. It indicates the numbers of misogynistic (Positive) and non-misogynistic (Negative) tweets in the training and test partitions.

Dataset	Training		Test		Total
	Positive	Negative	Positive	Negative	
Iber-SP	1649	1658	415	416	4138
Iber-En	1568	1683	283	443	3997
Eval-En	1785	2215	460	540	5000

The datasets of filtered song phrases resulting from the DA approach are described in Table 3 (they come from the sets of labeled song phrases introduced in Table 1). These song phrases were selected using the two proposed filtering

[8]https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestCentroid.html

[9]The phrase was obtained from the Spanish dataset, and it was slightly changed to respect the anonymity of users.

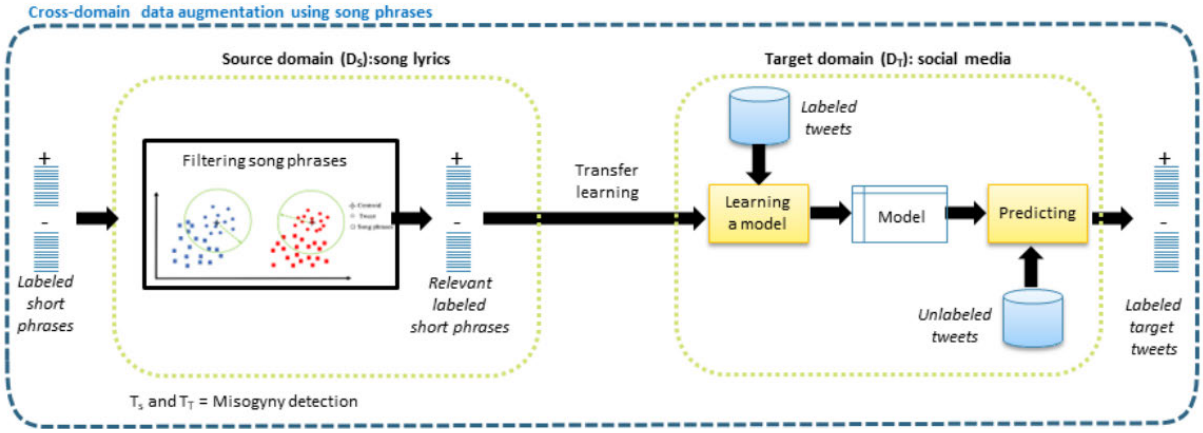


FIGURE 3. Overview of the cross-domain data-augmentation process using song phrases to support the classification of misogynistic tweets.

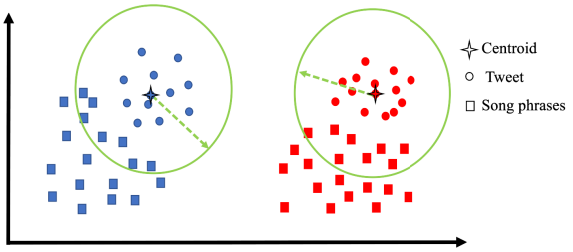


FIGURE 4. Representation of the filtering mechanism in the proposed data augmentation method. Red and blue elements represent the positive and negative classes, respectively.

TABLE 3. Distribution of the selected song phrases to augment the tweets from the target domain.

Dataset	Filter	Positive	Negative	Total
Spanish	Cosine	282	282	564
	Roccio	290	1411	1701
English	Cosine	424	424	848
	Roccio	1305	2095	3400

strategies; in the case of the cosine-based strategy, we employed a $\theta = 20\%$.[10]

B. TEXT REPRESENTATIONS

The tweets were lower-cased and tokenized into word unigrams. The special characters, emojis, URLs, and user mentions were removed. In addition, stopwords were deleted from the text for the BoW and Neural Networks based models, but they were kept for the case of BERT based models. To model the tweets, we applied the following text representations:

- *Bag of words (BoW).* A standard BoW using word unigrams weighted by normalized term frequency. This representation acts as the baseline method in the experiments.
- *Average Word Embeddings (AWE).* A vector-based representation built by averaging the embeddings of each tweet word. Specific word embeddings were trained on the misogynistic lyrics collection by using the *skip-gram* model[11] and FastText [61] through the *Genism* library[12] with *window_size* = 6. For comparison purposes, we also used general pre-trained embeddings,[13] which were learned with general information from *Wikipedia*. Specific and general embeddings are 300-dimensional vectors.
- *Neural Networks.* An attention-based GRU (*Gated Recurrent Unit*) network model. The (specific or general) word embeddings are used as the first layer in the model, which is sequentially connected to a BiGRU (*Bidirectional* GRU) with attention and a dense layer with *ReLU* (*Rectified Linear Unit*) activation.[14] The latter is introduced to a final dense sigmoid layer to generate the binary classification. We set the following parameters in training: *batch_size* = 32, *optimizer* = *Adam*, and *loss_function* = *binary_cross_entropy*.
- *Pretrained language models.* We used the pre-trained *distilbert-baseuncased* [62], a lighter version of BERT [20], for English experiments. Whereas for Spanish, we used BETO [63]. Both models were obtained from the *hugging-transformers* library.[15] To fine-tune the models, we used a *batch_size* = 16 and an early stopping mechanism.

[10]The thresholds of $\theta = 10\%$ and 30% were also tested. The best performance was obtained with 20%, suggesting that it represents a better trade-off between the quality of instances and the number of them.

[11]Both configurations, CBoW and Skip-gram, as well as Word2Vec were evaluated. The best results to detect misogyny were obtained with the Skip-gram FastText configuration.

[12]https://pypi.org/project/gensim/

[13]https://fasttext.cc/docs/en/pretrained-vectors.html

[14]We used a random search and 5-fold cross-validation to tune the number of units in these three layers, resulting in the following values: 256, 128, and 24, respectively.

[15]https://huggingface.co/docs/transformers/index

C. CLASSIFICATION AND EVALUATION

For classification, we applied different machine learning algorithms: a support vector machine (SVM), XGBoost (XG) [64], and Logistic Regression (LR). In line with the shared tasks using the same datasets, for the binary classification tasks, we mainly report the accuracy (Acc). However, for some experiments, we also show the F_1 scores. The experiments using neural and pre-trained models were run five times. In these cases, we report the average outcome and the standard deviation (std) on the test partition (Sections VI-B and VI-C).

VI. RESULTS AND DISCUSSION

This section reports the results that validate the ideas proposed in this research work. Section VI-A aims to evaluate conventional methods of transfer learning by leveraging song lyrics as a source domain to tackle the AMI task. The conducted experiments in Section VI-B evaluate the proposed DA approach. Finally, Section VI-C presents an analysis of different configurations of the approach and a comparison with state-of-the-art results.

A. EVALUATION OF TRANSFER LEARNING STRATEGIES

Two well-known transfer learning strategies were employed: domain adaptation and embedding-based methods. In the following, we report some experiments with these strategies.

1) DOMAIN ADAPTATION METHODS

In general, domain adaptation (DA) aims to train a classifier on one domain and test it on another with distinct data distribution [24], [65]. In this regard, the purpose of this experiment is to determine the relevance of transferring lyrics language as markers of misogyny in social media through a basic strategy of domain adaptation. Specifically, we train classifiers on song lyrics (i.e., source domain) to classify posts from social media datasets (i.e., target domain). Three classifiers were trained using different information from the source domain represented by a traditional BoW: first, the whole song lyrics (indicated as Lyrics); second, only the selected song phrases (SgPh), and third, tweets from social media together with song phrases (Tw ∪ SgPh). The trained models were applied to classify the unlabeled instances (i.e., the test data partitions) from the target domain into positive and negative categories. For comparison purposes, we also show the results from a BoW model trained and evaluated only with the data from the target domain (Tw). Table 4 shows the results obtained.

From the results, the following can be highlighted:

First, the use of song phrases was, in most cases, better than the use of whole songs, regardless of the classifier. This behavior was especially evident in the *Iber-Sp* and *Iber-En* collections. These results suggest that useful linguistic patterns for misogyny detection are concentrated in a few phrases of the songs, and not present across the entire songs.

TABLE 4. Results of a BOW model applying different domain adaptation configurations. Accuracy (Acc) and F_1 values are reported. The baseline classifier was trained only considering tweets from the target domain (Tw). Lyrics information was considered in training in three manners: whole song lyrics (Lyrics), song phrases (SgPh), and a combination of the latter with tweets (Tw ∪ SgPh).

Dataset	Train	SVM		XG		LR	
		Acc	F1	Acc	F1	Acc	F1
Iber-Sp	Tw	0.819	0.819	0.781	0.780	0.813	0.813
	Lyrics	0.659	0.646	0.656	0.648	0.651	0.629
	SgPh	0.665	0.656	0.657	0.649	0.649	0.641
	Tw ∪ SgPh	0.815	0.814	0.795	0.795	**0.824**	0.824
Iber-En	Tw	**0.806**	0.793	0.798	0.772	0.762	0.723
	Lyrics	0.665	0.591	0.675	0.574	0.678	0.601
	SgPh	0.686	0.633	0.669	0.574	0.697	0.654
	Tw ∪ SgPh	0.789	0.768	0.775	0.742	0.744	0.702
Eval-En	Tw	0.597	0.597	0.567	0.562	0.606	0.602
	Lyrics	0.638	0.634	0.642	0.627	**0.644**	0.636
	SgPh	0.615	0.614	0.641	0.633	0.623	0.622
	Tw ∪ SgPh	0.580	0.578	0.586	0.579	0.618	0.613

This finding can be exploited to enrich more robust DA approaches.

Second, although classifiers trained on song lyrics did not outperform the baseline in the first two collections, they show better results than a random classifier in binary scenarios, where the average results are close to 50%. Interestingly, in the *Eval-En* collection, these classifiers surpassed the baselines. This suggests the presence of a common subset of features across both domains that are valuable to detect misogyny, but different representations are necessary to explore in order to take advantage of them.

Third, the combination of feature spaces during training did not help the classification using the traditional BoW representation. Only a slight gain over the baseline was found with this combination in the Spanish dataset (around 1%).

In general, these findings motivated the development of the proposed approach to leverage song phrases as a data source to augment training sets in the target domain.

2) EMBEDDING-BASED METHODS

Word embeddings are another strategy for transfer learning that has contributed to many NLP (Natural Language Processing) tasks [66], [67], [68]. Hence, the main goal of this experiment is to evaluate embedding-based methods to transfer knowledge from song lyrics to the AMI task. In addition, we also aim to compare the use of specific against general embeddings. The former learned from song lyrics containing explicit misogyny content (i.e., they were trained only on song lyrics labeled as misogynistic), and the latter obtained from pre-trained word vectors from FastText. For this purpose, two basic embedding-based representations were evaluated: i) the average of the individual word embeddings (AWE). These vectors are classified through different machine learning algorithms, and ii) a GRU architecture fed by specific or general word embeddings. Their complementarity is also explored when both types of embeddings are combined as two input channels in the GRU (combination). In this experiment, the classification models are always trained and tested on the target domain. The GRU models

were run five times, and the average accuracy and std were reported. The results are shown in Table 5.

TABLE 5. Comparison of the performance (accuracy) of different classification models using specific and general embeddings, as well as a combination of them. Two text representations were evaluated, one based on average word embeddings (AWE) and the other learned by a GRU.

Dataset	Embeddings	AWE			GRU
		SVM	XG	LR	Avg ± std
Iber-Sp	General	0.776	0.781	0.762	0.779 ± 0.012
	Specific	**0.788**	0.771	0.777	**0.792 ± 0.018**
	Combination	-	-	-	0.782 ± 0.019
Iber-En	General	0.751	0.731	0.729	0.729 ± 0.029
	Specific	**0.792**	0.758	0.791	**0.742 ± 0.022**
	Combination	-	-	-	0.747 ± 0.039
Eval-En	General	0.641	0.625	**0.665**	0.576 ± 0.018
	Specific	0.615	0.617	0.618	**0.588 ± 0.026**
	Combination	-	-	-	0.586 ± 0.011

Through the experiments reported in the previous table, we note that: i) in most cases, domain-specific embeddings performed better than general ones, regardless of the model used (AWE or GRU). This indicates that song lyrics can generate valuable embeddings for the task at hand; ii) the combination of both types of embeddings by the GRU did not favor the construction of a robust classification model. These outcomes allow validating the knowledge transfer using specific embeddings into one downstream task.

B. EVALUATION OF THE PROPOSED APPROACH

Previous experiments have shown that knowledge from song lyrics can help detect misogynistic messages in social media. In this section, we evaluate the DA proposed method, which is aimed to increase the diversity of training data in the target domain (tweets) by adding examples from the source domain (song phrases). Since pre-trained linguistic models such as BERT have shown significant results in different text classification tasks, we fine-tuned BERT-base models to detect misogyny in social media. Specifically, we used *Distilbert* and *Beto* for English and Spanish datasets, respectively. For comparison purposes, we evaluated the same models but without data augmentation. Table 6 shows the results of these experiments.

From results, we can notice the following:

First, substantial improvements, according to the average values, were obtained when DA was performed over the baseline results (without DA), for all datasets and regardless of the filtering strategy used. Even in the case of DA, the minimum values were always better than those obtained without DA. In particular, the proposed approach outperformed the baseline results (No) with gains ranging from 1% in the *Iber-Sp* collection with a cosine-based filter, to 5.6% in the *Iber-En* dataset with a Roccio-based filter. In general, we consider that the augmented instances added new linguistic patterns to the training dataset in the target domain, therefore improving the models' performance.

TABLE 6. Performance (accuracy) of the proposed data augmentation technique using positive and negative song phrases selected by the cosine similarity and Roccio based filters. The model without data augmentation (No) is included as baseline.

Dataset	Augmentation	Average ± std	Minimum	Maximum
Iber-Sp	No	0.829 ± 0.015	0.810	0.854
	Cosine	0.839 ± 0.010	0.828	**0.857**
	Roccio	**0.844 ± 0.005**	0.835	0.852
Iber-En	No	0.836 ± 0.010	0.822	0.853
	Cosine	0.861 ± 0.019	0.835	0.888
	Roccio	**0.892 ± 0.009**	0.883	**0.906**
Eval-En	No	0.644 ± 0.018	0.617	0.671
	Cosine	**0.686 ± 0.016**	0.663	**0.705**
	Roccio	0.666 ± 0.006	0.659	0.677

Second, the gains obtained by the Roccio-based filter were higher than those from the *Cosine* mechanism in both *Iber-Sp* and *Iber-En* collections. However, in *Eval-En*, the *Cosine* filter yielded the best outcome. This suggests that the selection of one of them is dependent on the kind of dataset.

It should be noted that the proposed approach achieved better results than the conventional transfer learning strategies (which were evaluated in the previous section). Therefore, these results support the effectiveness of our method.

C. ANALYSIS

1) ANALYZING THE IMPACT OF DIFFERENT CONFIGURATIONS IN THE FILTERING MECHANISMS

In the previous experiments, we observed the relevance of evaluating the quality of song phrases through filters that discriminate noisy ones. In this section, we performed a deeper analysis of different settings that evaluate the instances in the source domain. First, we examined the effect produced by augmenting data with positive and negative song phrases, which were not filtered. This configuration is denoted as *All SgPh*. Second, we evaluated the DA with only positive (+) instances to emphasize the behavior of interest. Third, for comparison purposes, we show again the use of both positive and negative instances selected by the filters (these results correspond to *Cosine* and *Roccio* rows in the previous Table 6) and the baseline results when the models did not receive DA (No). Table 7 shows the results.

The following can be highlighted from the results: first, all DA configurations performed better than the baseline result that does not use augmentation (No), except *Cosine* (+) in *Iber-En*, even the cases where unfiltered song phrases were added (SgPh). These results corroborate the usefulness of augmenting data using song phrases. Second, the best result in each dataset was always obtained when the increased data were filtered with some configuration based on *Cosine* or *Roccio* mechanisms. The biggest differences with respect to baseline (No) correspond to: 2.2%, 5.6%, and 4.2% for *Iber-SP*, *Iber-En*, and *Eval-En* collections, respectively. This observation highlights the importance of adequately selecting song phrases to yield good performance. Third, using positive and negative instances works well, but adding only positive examples sometimes causes the same or even better results

TABLE 7. Evaluation of different settings in the filtering process; average, minimum, and maximum accuracy values are reported.

Dataset	Augmentation	Average ± std	Minimum	Maximum
Iber-Sp	No	0.829 ± 0.015	0.810	0.854
	All SgPh	0.841 ± 0.011	0.826	0.859
	Cosine (+)	**0.851 ± 0.005**	0.845	**0.859**
	Cosine (+) (-)	0.839 ± 0.010	0.828	0.857
	Roccio (+)	0.846 ± 0.004	0.846	0.856
	Roccio (+) (-)	0.844 ± 0.005	0.835	0.852
Iber-En	No	0.836 ± 0.010	0.822	0.853
	All SgPh	0.860 ± 0.016	0.844	0.886
	Cosine (+)	0.825 ± 0.013	0.810	0.842
	Cosine (+) (-)	0.861 ± 0.019	0.835	0.888
	Roccio (+)	0.843 ± 0.030	0.803	0.868
	Roccio (+) (-)	**0.892 ± 0.009**	0.883	**0.906**
Eval-En	No	0.644 ± 0.018	0.617	0.671
	All SgPh	0.684 ± 0.010	0.666	0.694
	Cosine (+)	0.682 ± 0.016	0.658	0.705
	Cosine (+) (-)	**0.686 ± 0.016**	0.663	**0.705**
	Roccio (+)	0.652 ± 0.004	0.645	0.656
	Roccio (+) (-)	0.666 ± 0.006	0.659	0.677

(e.g., in the *Iber-SP* dataset). Therefore, it is convenient to further explore these configurations in each dataset.

In general, the results obtained confirm that song phrases can help increase the training data and improve the classifiers' performance, but also highlight the fundamental role of the filtering mechanisms to select the most pertinent song phrases.

2) COMPARISON WITH STATE-OF-THE-ART RESULTS

This section compares the proposed approach with state-of-the-art methods with the intention of assessing its competitiveness. Figure 5 presents the distribution of the official results by the shared tasks participants in the same datasets as those used in our experiments. In the figure, we indicate with red crosses our best results, using the best configuration for each dataset as previously reported in Table 6. Particularly, they correspond to Bert-base models using the following settings in the filtering mechanism: *Cosine* (+), *Roccio* (+)(−), and *Cosine* (+)(−) for *Iber-Sp*, *Iber-En*, and *Eval-En*, respectively.

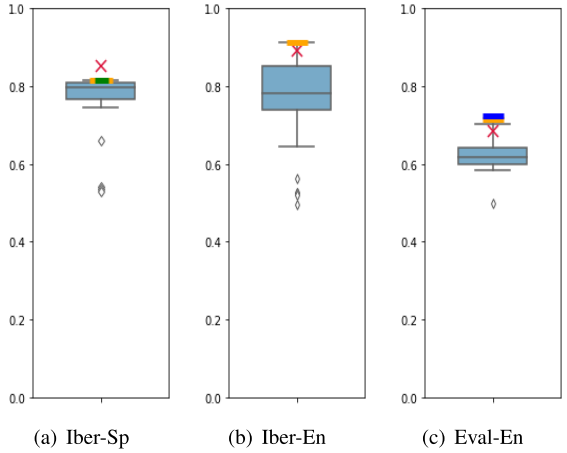


FIGURE 5. Distribution of results (in accuracy) of the participants of the shared tasks a) IberEval-Spanish, b) IberEval-English, and c) EVALITA-English. The red crosses indicate the performance of our best models in each collection.

From Figure 5, it is possible to appreciate competitive results with respect to the participating teams. A special case is the result in the Spanish collection of *IberEval*, where our approach obtained better results than the winner of the shared task. On the other hand, in the *IberEval* English dataset, the performance obtained would have placed our approach in the fourth position. In the *Evalita* shared task, the results obtained are slightly lower than those of the winning team, and would place the proposed approach in the second position. At this point, it is important to mention that although the winners of *IberEval* used basic BOW-based representations, their systems were enriched with other elements beyond words, such as hashtags, emojis, links, and others. Moreover, in *Evalita*, the winner method concatenated different features: sentence embeddings, TF-IDF (*Term Frequency-Inverse Document Frequency*) vectors, and BOW vectors. In contrast, the proposed approach only exploits the words contained in the texts through BERT-base models, since the idea was to evaluate the potential of learning transfer between the mentioned domains by applying data augmentation.

Other authors have also used the same datasets for their research, obtaining in some cases very interesting results. For example, [26] describes some approaches based on cross-domain classification using several datasets of other abusive language phenomena (e.g., sexism, hate speech, and offensive language). These approaches included hashtags and emojis in the text representation and used different classifiers such as SVM, GRU, and BERT. Their results are indicated in Figure 5 with orange hyphens and correspond to the following accuracy values: 91.32, 81.47, 71.6 for *Iber-En*, *Iber-Sp* and *Eval-En* datasets, respectively. On the other hand, the approach in [19], represented with a blue hyphen in the figure (*accuracy* = 0.724 in the EVALITA English dataset), used an LSTM[16]-based Bayesian transfer learning method, which was pre-trained using three datasets, corresponding to general and task-specific domains. Finally, the method reported in [69] used a combination of embeddings and linguistic features; its result in the IberEval Spanish collection (*accuracy* = 0.815) is indicated with a green marker in the figure.

To conclude, the proposed method performed better in Spanish than in English, but in all cases, showed competitive results compared to the robust methods mentioned above.

VII. CONCLUSION

Inspired by the frequent prevalence of misogynistic ideology exposed in music, which reflects socio-cultural beliefs, this paper explored the relevance of song lyrics content as a data source for modeling verbal misogyny and transferring knowledge to the task of Automatic Misogyny Identification in social media. In particular, we proposed a data augmentation

[16] *Long short-term memory.*

approach that leverages song phrases to increase training data in the target domain (social media posts). The main idea behind this approach is to improve the generalization ability of learning models by transferring the semantic richness of song lyrics, which reflect the social context. Besides, we introduced a methodology to build a labeled dataset with relevant song phrases where the positive class is sensitive to misogyny.

The proposed approach was evaluated through an experimental study on benchmark collections containing English and Spanish tweets. The results were encouraging since they outperformed those of state-of-the-art in the Spanish collection and were competitive in English, which were obtained by methods computationally more expensive than ours. These results confirmed the suitability of the approach for the task at hand. In particular, we obtained the following conclusions:

- The song lyrics contain valuable information that can be transferred to the AMI task. However, the relevant features are concentrated only on some phrases and not on the entire songs. Therefore, their appropriate selection is important to be used by transfer approaches.
- The proposed approach is suitable for transferring knowledge between the aforementioned domains. Its performance is better than other conventional transfer learning models, such as domain adaptation and embedding-based methods.
- A proper quality assessment of the augmented song phrases is essential for the proposed approach. The added phrases capture new linguistic patterns that improve the generalization ability of the models. In general, we believe that the increasing availability and diversity of music make it a valuable source of information for transferring knowledge.

We hope this work provides a general framework for future research on the richness of song lyrics for transferring knowledge into other domains and tasks. Our findings suggest opportunities for future research in modeling other behaviors, such as aggression and bullying. Furthermore, given that these conditions are common worldwide and transcend linguistic boundaries, we plan to evaluate the proposed approach in other languages, such as Italian and French. Finally, we also plan to study its contribution to multimodal tasks.

REFERENCES

[1] C. Napp and T. Breda, "The stereotype that girls lack talent: A worldwide investigation," *Sci. Adv.*, vol. 8, no. 10, Mar. 2022, Art. no. eabm3689.

[2] K. Srivastava, S. Chaudhury, P. S. Bhat, and S. Sahu, "Misogyny, feminism, and sexual harassment," *Ind. Psychiatry J.*, vol. 26, no. 2, pp. 111–113, 2017.

[3] S. Kumar, P. Tyagi, S. Saxena, and S. Badu, "Right to equality and gender justice," *J. Positive School Psychol.*, vol. 6, pp. 4916–4925, May 2022.

[4] P. Casares, "Discurso de odio y feminicidios en México," *Tram[p]as de la Comunicación y la Cultura*, no. 66, pp. 29–35, Apr. 2009.

[5] K. Barker and O. Jurasz, "Online violence against women as an obstacle to gender equality: A critical view from Europe," *Eur. Equality Law Rev.*, vol. 2020, no. 1, pp. 47–60, 2020.

[6] E. A. Jane, *Misogyny Online: A Short (and Brutish) History*. Newbury Park, CA, USA: Sage, 2016.

[7] E. Fersini, P. Rosso, and M. Anzovino, "Overview of the task on automatic misogyny identification at IberEval 2018," in *Proc. CEUR Workshop Proc.*, vol. 2150, 2018, pp. 214–228.

[8] M. Anzovino, E. Fersini, and P. Rosso, "Automatic identification and classification of misogynistic language on Twitter," in *Proc. Int. Conf. Appl. Natural Lang. Process. Inf. Syst.* Cham, Switzerland: Springer, 2018, pp. 57–64.

[9] V. Basile, M. D. Maro, D. Croce, and L. Passaro, "EVALITA 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for Italian," in *Proc. 7th Eval. Campaign Natural Lang. Process. Speech Tools Italian. Final Workshop, (EVALITA)*, vol. 2765, 2020, pp. 1–7.

[10] R. Tsokalidou, "Linguistic misogyny—A language universal: Observations, questions and ideas," *Sel. Papers Theor. Appl. Linguistics*, vol. 3, pp. 363–381, Jun. 1989.

[11] D. Defranza, H. Mishra, and A. Mishra, "How language shapes prejudice against women: An examination across 45 world languages," *J. Personality Social Psychol.*, vol. 119, no. 1, p. 7, 2020.

[12] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 78–84.

[13] P. Zeinert, N. Inie, and L. Derczynski, "Annotating online misogyny," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, 2021, pp. 3181–3197.

[14] E. Shushkevich and J. Cardiff, "Automatic misogyny detection in social media: A survey," *Computación Y Sistemas*, vol. 23, no. 4, pp. 1159–1164, 2019.

[15] E. Aldana-Bobadilla, A. Molina-Villegas, Y. Montelongo-Padilla, I. Lopez-Arevalo, and O. S. Sordia, "A language model for misogyny detection in Latin American Spanish driven by multisource feature extraction and transformers," *Appl. Sci.*, vol. 11, no. 21, Nov. 2021, Art. no. 10467.

[16] T. Farrell, M. Fernandez, J. Novotny, and H. Alani, "Exploring misogyny across the manosphere in reddit," in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 87–96.

[17] R. Fulper, G. L. Ciampaglia, E. Ferrara, F. Menczer, Y.-Y. Ahn, A. Flammini, B. Lewis, and K. Rowe, "Misogynistic language on Twitter and sexual violence," in *Proc. ACM Web Sci. Workshop Comput. Approaches Social Modeling (ChASM)*, 2014, pp. 6–9.

[18] S. Hewitt, T. Tiropanis, and C. Bokhove, "The problem of identifying misogynist language on Twitter (and other online social spaces)," in *Proc. 8th ACM Conf. Web Sci.*, May 2016, pp. 333–335.

[19] M. A. Bashar, R. Nayak, and N. Suzor, "Regularising LSTM classifier by transfer learning for detecting misogynistic tweets with small training set," *Knowl. Inf. Syst.*, vol. 62, no. 10, pp. 4029–4054, Oct. 2020.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[21] E. Fersini, D. Nozza, and P. Rosso, "Overview of the EVALITA 2018 task on automatic misogyny identification (AMI)," in *Proc. 6th Eval. campaign Natural Lang. Process. Speech tools Italian (EVALITA'18)*, T. Caselli, N. Novielli, V. Patti, and P. Rosso, Eds. Turin, Italy, 2018, pp. 1–9.

[22] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 54–63.

[23] P. R. E. Fersini and D. Nozza, "AMI @ EVALITA2020: Automatic misogyny identification," in *Proc. 7th Eval. Campaign Natural Lang. Process. Speech tools Italian (EVALITA 2020)*, V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, Eds., 2020.

[24] Z. Alyafeai, M. S. AlShaibani, and I. Ahmad, "A survey on transfer learning in natural language processing," 2020, *arXiv:2007.04239*.

[25] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, Dec. 2016.

[26] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny detection in Twitter: A multilingual and cross-domain study," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102360.

[27] R. E. Ramos-Vargas, I. Román-Godínez, and S. Torres-Ramos, "Comparing general and specialized word embeddings for biomedical named entity recognition," *PeerJ Comput. Sci.*, vol. 7, p. e384, Feb. 2021.

[28] I. Goenaga, A. Atutxa, K. Gojenola, A. Casillas, A. D. D. Ilarraza, N. Ezeiza, M. Oronoz, A. Pérez, and O. P.-D. V. Naspre, "Automatic misogyny identification using neural networks," in *Proc. IberEval@ SEPLN*, 2018, pp. 249–254.

[29] S. Fabrizi, "Fabsam@ AMI: A convolutional neural network approach," in *Proc. 7th Eval. Campaign Natural Lang. Process. Speech Tools Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR, 2020, pp. 1–5.

[30] S. Dutta, U. Majumder, and S. K. Naskar, "An efficient bert based approach to detect aggression and misogyny," in *Proc. 18th Int. Conf. Natural Lang. Process. (ICON)*, 2021, pp. 493–498.

[31] N. S. Samghabadi, P. Patwa, S. Pykl, P. Mukherjee, A. Das, and T. Solorio, "Aggression and misogyny detection using bert: A multi-task approach," in *Proc. 2nd Workshop Trolling, Aggression Cyberbullying*, 2020, pp. 126–131.

[32] H. T. Ta, A. B. S. Rahman, L. Najjar, and A. Gelbukh, "Transfer learning from multilingual deberta for sexism identification," in *Proc. CEUR Workshop*, vol. 3202, 2022, pp. 1–10.

[33] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, Jan. 2022.

[34] D. Cooke, *The Language of Music*. London, U.K.: Oxford Univ. Press, 1959.

[35] J. Bicknell, "Can music convey semantic content? A Kantian approach," *J. Aesthetics Art Criticism*, vol. 60, no. 3, pp. 253–261, Aug. 2002.

[36] G. R. Fischer, "How music communicates," *Semiotica*, vol. 53, nos. 1–3, pp. 1–3, 1985.

[37] G. Barton, *The Relationship Between Music, Culture, and Society: Meaning in Music*. Cham, Switzerland: Springer, 2018.

[38] D. Edmonds and J. A. Sedoc, "Multi-emotion classification for song lyrics," in *Proc. 11th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, Apr. 2021, pp. 221–235.

[39] K. Matsumoto and M. Sasayama, "Lyric emotion estimation using word embedding learned from lyric corpus," in *Proc. IEEE 4th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2018, pp. 2295–2301.

[40] M. Rospocher, "Explicit song lyrics detection with subword-enriched word embeddings," *Expert Syst. Appl.*, vol. 163, Jan. 2021, Art. no. 113749.

[41] M. McVicar, B. D. Giorgi, B. Dundar, and M. Mauch, "Lyric document embeddings for music tagging," in *Proc. CMMR*, 2021, p. 47.

[42] M. P. Barman, A. Awekar, and S. Kothari, "Decoding the style and bias of song lyrics," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 1165–1168.

[43] M. Aghzal and A. Mourhir, "Distributional word representations for code-mixed text in Moroccan darija," *Proc. Comput. Sci.*, vol. 189, pp. 266–273, Jan. 2021.

[44] P. Freudiger and E. M. Almquist, "Male and female roles in the lyrics of three genres of contemporary music," *Sex Roles*, vol. 4, no. 1, pp. 51–65, Feb. 1978.

[45] C. M. Frisby and E. Behm-Morawitz, "Undressing the words: Prevalence of profanity, misogyny, violence, and gender role references in popular music from 2006–2016," *Media Watch*, vol. 10, no. 1, pp. 5–21, Jan. 2019.

[46] S. de Boise, "Music and misogyny: A content analysis of misogynistic, antifeminist forums," *Popular Music*, vol. 39, nos. 3–4, pp. 459–481, Dec. 2020.

[47] R. M. Gourdine and B. P. Lemmons, "Perceptions of misogyny in hip hop and rap: What do the youths think?" *J. Hum. Behav. Social Environ.*, vol. 21, no. 1, pp. 57–72, Jan. 2011.

[48] E. Tobias, "Flipping the misogynist script: Gender, agency, hip hop and music education," *Action, Criticism Theory Music Educ.*, vol. 13, no. 2, pp. 1–37, 2014.

[49] G. Rebollo-Gil and A. Moras, "Black women and black men in hip hop music: Misogyny, violence and the negotiation of (white-owned) space," *J. Popular Culture*, vol. 45, no. 1, pp. 118–132, Feb. 2012.

[50] R. Weitzer and C. E. Kubrin, "Misogyny in rap music: A content analysis of prevalence and meanings," *Men Masculinities*, vol. 12, no. 1, pp. 3–29, Oct. 2009.

[51] G. Cundiff, "The influence of rap and hip-hop music: An analysis on audience perceptions of misogynistic lyrics," *Elon J. Undergraduate Res. Commun.*, vol. 4, no. 1, pp. 3–4, 2013.

[52] J. Q. K. Ling and F. Genevieve Dipolog-Ubanan, "Misogyny in the lyrics of billboard's top rap airplay artists," *Int. J. Arts Humanities Social Sci.*, vol. 2, no. 6, pp. 7–13, 2017.

[53] C. d'Hont, "How female is the future? Undoing sexism in contemporary metal music," in *Misogyny, Toxic Masculinity, Heteronormativity Post-2000 Popular Music*. Cham, Switzerland: Springer, 2021, pp. 95–112.

[54] M. A. Flynn, C. M. Craig, C. N. Anderson, and K. J. Holody, "Objectification in popular music lyrics: An examination of gender and genre differences," *Sex Roles*, vol. 75, nos. 3–4, pp. 164–176, Aug. 2016.

[55] F. Kleedorfer, P. Knees, and T. Pohle, "Oh oh oh whoah! towards automatic topic detection in song lyrics," in *Proc. Ismir*, Jan. 2008, pp. 287–292.

[56] B. Bretthauer, T. S. Zimmerman, and J. H. Banning, "A feminist analysis of popular music," *J. Feminist Family Therapy*, vol. 18, no. 4, pp. 29–51, Feb. 2007.

[57] T. M. Adams and D. B. Fuller, "The words have changed but the ideology remains the same: Misogynistic lyrics in rap music," *J. Black Stud.*, vol. 36, no. 6, pp. 938–957, Jul. 2006.

[58] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Detecting misogyny and xenophobia in Spanish tweets using language technologies," *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–19, May 2020.

[59] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Jan. 2021.

[60] W. Pan, E. Zhong, and Q. Yang, "Transfer learning for text mining," in *Mining Text Data*. Cham, Switzerland: Springer, 2012, pp. 223–257.

[61] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Jun. 2016.

[62] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[63] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," in *Proc. pML4DC AT ICLR*, 2020, pp. 1–10.

[64] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[65] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems*, vol. 19, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA, USA: MIT Press, 2006, pp. 1–8.

[66] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proc. Workshop ICLR*, Jan. 2013, pp. 1–12.

[67] F. Almeida and G. Xexéo, "Word embeddings: A survey," 2019, *arXiv:1901.09069*.

[68] S. S. Birunda and R. K. Devi, "A review on word embedding techniques for text classification," in *Innovative Data Communication Technologies and Application*, J. S. Raj, A. M. Iliyasu, R. Bestak, and Z. A. Baig, Eds. Singapore: Springer, 2021, pp. 267–281.

[69] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," *Future Gener. Comput. Syst.*, vol. 114, pp. 506–518, Jan. 2021.

RICARDO CALDERÓN-SUAREZ received the M.S. degree in optical computing from the Polytechnic University of Tulancingo, Hidalgo, Mexico, in 2018, where he is currently pursuing the Ph.D. degree in optomechatronics. His research interests include natural language processing, machine learning, and computer vision.

ROSA M. ORTEGA-MENDOZA received the Ph.D. degree in computer science from the Autonomous University of Hidalgo, Mexico. From 2018 to 2019, she was a Postdoctoral Researcher at the Language Technologies Group, National Institute of Astrophysics, Optics and Electronics (INAOE). She is a Research Professor with the Graduate Studies Department, Polytechnic University of Tulancingo, Hidalgo, Mexico. Her research interests include natural language processing, text mining, authorship analysis, social media data analysis, and machine learning.

MANUEL MONTES-Y-GÓMEZ received the Ph.D. degree in computer science from the Computing Research Center, National Polytechnic Institute, Mexico. He is a Researcher with the National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico. He is a Visiting Professor with the Polytechnic University of Valencia, the University of Geneva, and the University of Alabama at Birmingham. He is an author of more than 250 journals and conference papers in the fields of information retrieval, text mining, and authorship analysis. His research interest includes automatic text processing. He is a Founding Member of the Mexican Association of Natural Language Processing and an Organizer of the National Workshop on Language Technologies, the Mexican Workshop on Plagiarism Detection and Authorship Analysis, and the Mexican Autumn School on Language Technologies.

CARINA TOXQUI-QUITL received the B.S. degree from the Puebla Autonomous University, Mexico, in 2004, and the M.S. and Ph.D. degrees in optics from the National Institute of Astrophysics, Optics and Electronics, in 2006 and 2010, respectively. She is an Assistant Professor with the Polytechnic University of Tulancingo. Her current research interests include invariant feature extraction, classification, and computer vision.

MARCO A. MÁRQUEZ-VERA received the Ph.D. degree in computer science from the Autonomous University of Hidalgo, Mexico. He is a Professor with the Polytechnic University of Pachuca, Hidalgo, Mexico. His research interests include deep learning, swarm intelligence, and fuzzy logic. He is a member of the National System of Researchers Level 1 by CONACyT, Mexico. He is an Associate Editor of the journals, such as *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika* and *International Journal of Robotics and Control Systems*.