IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 17, 2020, accepted November 30, 2020, date of publication December 4, 2020,
date of current version December 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3042604

Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data

FRANCISCO RODRÍGUEZ-SÁNCHEZ, JORGE CARRILLO-DE-ALBORNOZ, AND LAURA PLAZA
NLP & IR Group, UNED, 28040 Madrid, Spain

Corresponding author: Francisco Rodríguez-Sánchez (frodriguez.sanchez@invi.uned.es)

This work was supported by the Spanish Ministry of Science and Innovation under Project Misinformation and Miscommunication in Social Media (PGC2018-096212-B-C32).

ABSTRACT During the last decade, hateful and sexist content towards women is being increasingly spread on social networks. The exposure to sexist speech has serious consequences to women's life and limits their freedom of speech. Previous studies have focused on identifying hatred or violence towards women. However, sexism is expressed in very different forms: it includes subtle stereotypes and attitudes that, although frequently unnoticed, are extremely harmful for both women and society. In this work, we propose a new task that aims to understand and analyze how sexism, from explicit hate or violence to subtle expressions, is expressed in online conversations. To this end, we have developed and released the first dataset of sexist expressions and attitudes in Twitter in Spanish (MeTwo) and investigate the feasibility of using machine learning techniques (both traditional and novel deep learning models) for automatically detecting different types of sexist behaviours. Our results show that sexism is frequently found in many forms in social networks, that it includes a wide range of behaviours, and that it is possible to detect them using deep learning approaches. We discuss the performance of automatic classification methods to deal with different types of sexism and the generalizability of our task to other subdomains, such as misogyny.

INDEX TERMS Sexism detection, social media, natural language processing, machine learning.

I. INTRODUCTION

The rapid development of web technologies and social networks has enabled the interaction between people from different countries, cultures and ethnicities. Although the advantages and positive effects of this global communication are obvious, the invisibility, anonymity and accessibility have made the expression of xenophobic, racist and sexist discourses easy and unpunished. The so-called *online dis-inhibition effect* [1] emboldens users to engage in behaviors they are unlikely to perform face-to-face. Moreover, the quick spread of online information, especially in social networks, has made these harassment behaviours extremely dangerous, so that solutions are required to effectively reduce the harm caused by hateful propaganda in cyberspace.

Hate speech can be defined as *language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group* [2]. Internet intermediaries like Facebook have recently announced increasing efforts to moderate and fight against hateful and harmful speech [3], so that they can protect their users from harassment and hateful language. However, they recognized to be failing to detect some content of this kind [4]. Other companies, like Twitter, are continuously reviewing their policies to include additional types of abusive behavior and creating new ways to eradicate hateful content from their websites, that range from warnings to the deletion of harmful tweets and even to the permanent suspension of users.[1] However, despite making a great effort and using many human resources, they are facing many difficulties when dealing with the huge amount of data generated by users [5].

During the last years, the role of women within online platforms has gained attention, unfortunately because of the growing hatred and abuse against them. According to the work of [6], women are about twice as likely as men to say they have been harassed online as a result of their gender.

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-keung Fung.

[1] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Recently, Amnesty International published a report[2] where they describe Twitter as a ''toxic place'' for women. According to this report, Twitter is promoting violence and hate against or threaten people based on their gender. The report also suggests that Twitter is failing to protect women against harassment and it could have a negative impact on their freedom of speech. Moreover, previous works have found a relationship between verbal harassment to woman and action. Fulper *et al.* [7], for instance, demonstrated the existence of a correlation between the number of rapes and the amount of misogynistic tweets per state in USA. Therefore, fighting against online sexism is a social urgency.

The Oxford English Dictionary defines **sexism** as *prejudice, stereotyping or discrimination, typically against women, on the basis of sex*.[3] Similarly, the Real Academia Española de la Lengua defines it as *discrimination of people on the basis of sex*.[4] In general, sexist behaviours and discourses underestimate the role of women. Inequality and discrimination against women that remains embedded in society is increasingly being replicated online [8]. The internet perpetuates and even naturalizes gender differences and sexist attitudes [12]. Moreover, given that an important percentage of Internet users (especially social networks users) are teenagers, the increasing sexism on the Internet requires urgent study and social debate that leads to actions. However, detecting online sexism may be difficult, as it may be expressed in very different forms. Sexism may sound ''friendly'': the statement ''*Women must be loved and respected, always treat them like a fragile glass*'' may seem positive, but is actually considering that women are weaker than men. Sexism may sound ''funny'', as it is the case of sexist jokes or humour (''*You have to love women...just that...You will never understand them.*''). Sexism may sound ''offensive'' and ''hateful'', as in ''*Humiliate, expose and degrade yourself as the fucking bitch you are if you want a real man to give you attention*''. However, even the most subtle forms of sexism can be as pernicious as the most violent ones and affect women in many facets of their lives [10], [11], including domestic and parenting roles, career opportunities, sexual image and life expectations, to name a few.

Current research on sexism in online media is focused on detecting misogyny or hatred towards women. The Oxford English Dictionary defines **misogyny** as *hatred or dislike of, or prejudice against women*. Sexism does not always imply misogyny: when a man claims that he prefers his wife to stay at home because this way she can attend his children better, he is being sexist. When a man claims that wives should only be allowed to stay at home, he is being misogynist. Both attitudes are supporting a stereotype against a woman, but the second expresses hostility and prejudice against her. Therefore, works dealing with the detection of misogyny are forgetting a wide spectrum of sexist attitudes and behaviours that are, in fact, the most frequent and dangerous for the society. Our aim is the detection of sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviours. To the best of our knowledge, no previous work has addressed the detection of this implicit, and not necessarily violent, sexism in social network conversations.

In this paper, we aim to understand how sexist behaviours, beliefs and attitudes are expressed in Twitter conversations. We focus on tweets written in Spanish, although the method employed and the conclusions extracted are directly applicable to other languages. We propose to identify sexism in social networks using an automatic system based on machine learning. First, we collect tweets automatically and compose a new dataset, MeTwo. Then, a series of machine learning algorithms are applied to classify the tweets into sexist or non-sexist. Finally, an exhaustive analysis is performed to study the properties of our system. This is a first step towards the more ambitious goal of creating novel mechanisms to detect and alert from abusive and sexist behaviours against women in social media.

Our results show that we can successfully detect different types of sexism. We prove that misogyny and hatred towards women are easier to detect than subtle or non-hateful sexism since it is much less context dependent. However, we observe that subtle sexism is expected to be more frequently found in our corpus and includes itself a wide range of behaviours. We discuss the performance of automatic classification methods to deal with all these types of sexism. Similarly, we show that a classification system trained on our dataset is able to generalize better than the same system trained on a dataset of misogynistic expressions.

In summary, our work has the following contributions:

- A new task is proposed in the area of offensive language detection that broaden the scope of modeling sexism detection in social media. We consider sexist attitudes that affect different facets of women and that may be subtle or hostile.
- The construction and manual annotation of the first Spanish corpus of sexist expressions in Twitter (the MeTwo dataset) that may be employed by the research community to advance the state of the art in the proposed task.
- The development of machine learning methods to automatically detect sexism in tweets, including the comparison of traditional ML methods with novel approaches based on neural networks and transfer learning. We achieve considerable results outperforming the baselines proposed and discuss the implications and generalization of our approach.

The rest of this paper is organized as follows: in section 2, we discuss related works. In section 3, the annotated corpus is presented. In section 4, we describe the classification system. Results and analysis are presented in section 5.

[2]https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/

[3]https://www.oed.com/

[4]http://dle.rae.es/srv/search?m=30&w=sexismo

Finally, the conclusions and future works are given in section 6.

II. RELATED WORK

Substantial work has been devoted to the detection of hate speech in recent years, including tasks such as racist or xenophobic content detection, but few works have faced sexism detection and, in particular, they have dealt with sexism as the detection of hate speech against women. Consequently, they have worked with hostile and explicit sexism, overlooking subtle or implicit expressions of sexism. However, some ideas and techniques from hate speech detection may apply to our problem. Therefore, in this section, we briefly review related work in the hate speech field along with previous works on sexism and misogyny detection.

A. HATE SPEECH DETECTION

The hate speech detection task consists in detecting hateful content in online communities. With the spread of Internet and the growth of online interactions, many users have the risk of being harassed on social media, blogs or forums. This particular form of harassment can have a negative impact on their online experience and the community in general. Sentences such as *"this niggers are taking the jobs of locals"* or *"Every human being who is a #muslim should be killed. End of story. #islam #IslamicState"* targeting minority groups are seen online daily.

In recent years, the Artificial Intelligence for Social Good (AI4SG) movement has promoted the creation of applications to protect minority groups on the Internet [16]. Some works have dealt with the prevention of sexual harassment [17], sexual discrimination detection [18], and cyberbully and trolling [19]. More recent works have dealt with suicidal ideation detection to address some real consequences of hate on the Internet [20].

First works on hate speech detection were based on bag-of-words (BOW) approaches [2], [14], [15]. In 2012, we find one of the earlier researches using machine learning based classifiers for detecting abusive language [21] as opposed to the pattern-based methods [22]. Traditional machine learning algorithms such as logistic regression [14], support vector machines and decision trees [2] have been widely employed to detect hate speech. Non-linguistic features like the gender or ethnicity of the author can help improve hate speech classification but this information is often unavailable or unreliable on social media [15]. There are also approaches that include some forms of sentiment information as features. Hate is a negative emotion, and thus it is fair to suppose that messages expressive negative emotions are more probably expressing hate than those expressing neutral o positive emotions. As a result, sentiment analysis and polarity detection techniques are usually applied to hate detection [13], [23].

Also inspired by the sentiment analysis and affective computing works, the use of external lexical resources has been applied to hate speech detection [24]. In [25], a lexicon of hate verbs which condone or encourage acts of violence is developed. Since lexicon-based classification depends greatly on the availability of high quality external resources, other works merge the benefits of the machine learning and lexicon-based classification approaches to detect hate speech [26].

During the last years, the application of neural models to hate speech detection has gained attention. These models typically apply deep learning approaches such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) [27]–[30], showing impressive results in many tasks related to natural language processing [31]. In this paper, we apply some of these methods to our problem along with transfer learning techniques [32].

Due to the availability of resources, the majority of studies on hate speech detection works on English texts. However, the academic event SemEval 2019 (Task 5) aimed to detect hate speech against immigrants and women in Spanish and English messages extracted from Twitter [33] and promoted research in Spanish. In this task, deep learning approaches were proposed to detect misogyny and racism in texts in Spanish [34]. Similarly, [37] aims to detect cyber hate speech in Arabic tweets employing a wide range of traditional machine learning techniques.

B. MISOGYNY DETECTION

"Misogyny" and "sexism" are frequently considered interchangeable, though both terms have different nuances. Currently, the definition of misogyny is under discussion [35], [36]. However, the most widely accepted definition of misogyny implies the expression of hostility and hatred towards women. In contrast, sexism comprises any form of oppression or prejudice against women and therefore may be hostile (as in the case of misogyny) or subtle. Thus, sexism includes misogyny but is not limited to it.

Current studies on the identification of sexism are related to hate speech detection. One of the first datasets was developed to study sexism in conjunction with racism [14], [15] . However, this dataset only comprises the expression of hate or hostile sexism towards women, overlooking other kinds of sexism. Sharifirad and Jacovi [38] presented a categorization of sexism that included indirect, sexual, and physical sexism. A more recent study by [39] seeks to categorize accounts of sexism. Because the growing interest of hate detection towards women, other tasks to protect women from hate on the internet have emerged. For instance, sexist MEME detection [40] and classification of sexist advertisements [41].

We can find in the literature previous works that have specifically faced the automatic detection of misogyny in text [42]–[45] as well as some datasets annotated with misogynist expressions [33]. ElSherief *et al.* [46] compiled Hate Lingo, an English dataset that comprises hate speech tweets that include hatred expressions towards people based on some intrinsic characteristics of the person, including their gender, class, ethnicity or religion. Similarly, Ousidhoum *et al.* [47] create a multi-lingual corpus that included expressions of hate towards women in English, French and Arabic. Similarly,

El Ansari *et al.* [48] construct a dataset combining manual and automatic annotation of Arabic texts addressing discrimination and violence against women.

Recently, the IberEval competition focused on the automatic identification of misogyny in Twitter [49]. Teams were proposed to identify misogynist tweets both in Spanish and English. The corpus created in this competition (AMI dataset) was also used to build the dataset for SemEval 2019 Task 5 [33]. Approaches presented to the competition were mainly based on supervised machine learning on different textual features (such as unigrams and bigrams, sentiment-based information, or syntactic categories) or user-based features (such as the number of retweets, followers, etc.) [50]–[52]. The use of lexical resources for extracting signals (such as swear word count, sexist slurs presence) showed excellent performance in the task [53]. Deep learning methods, such as recurrent neural networks, are explored in [54] along with word embedding features.

However, to the best of our knowledge, no previous work has explicitly tackled subtle sexism in social networks, nor has produced annotated datasets that enable the study and detection of the broad spectrum of behaviors and expressions that sexism encompasses. In this work, we develop the MeTwo dataset, a corpus of sexist tweets in Spanish that aims to help automatic systems to detect the broad spectrum of sexist attitudes that occur in Twitter. The dataset is accompanied by an exhaustive study and categorization of frequent sexist expressions in social networks.

III. MeTwo: MACHISMO AND SEXISM TWITTER IDENTIFICATION DATASET

In this section, the MeTwo dataset, a corpus for the detection of sexist expressions and attitudes in Twitter, is presented. The method to compile and annotate the corpus is described and some data statistics are shown. As already mentioned, although some datasets of misogyny and hate speech are available [15], [33], [44], to the best of our knowledge, MeTwo is the first corpus in Spanish designed to identify sexism in a broad sense, from hostile to much more subtle sexism. The MeTwo dataset is available for research at Github.[5]

A. CORPUS COLLECTION

To bootstrap our dataset, we first collected a number of popular expressions and terms commonly used to underestimate the role of women in our society, encourage the harassment towards them or limit their freedom of speech. Our main source for such expressions was the Twitter account of the Spanish journalist Ana Isabel Bernal-Triviño,[6] which collects phrases and expressions that women (Twitter users) have received on a day-to-day basis, and that have made them feel belittled and undermined because of their genre. We manually inspected them to select both expressions that may be clearly offensive, or even violent, and expressions that are subtle or even normalized. Table 1 shows the terms and expressions selected to build the MeTwo dataset. Initially, a total of 29 Spanish terms and expressions were considered as keywords to create the corpus.

TABLE 1. Tweets collected per term.

	Term	Term English	No. Tweets
1	como una mujer	as a woman	15094
2	feminazi	feminazi	15093
3	a la cocina	go to the kitchen	15087
4	zorra	slut	15086
5	loca del	crazy about	15084
6	como una nina	as a girl	15080
7	las feministas	the feminist	15076
8	niñata	little girl	15032
9	en tus dias	in your days	14190
10	a fregar	go washing	14013
11	mojigata	puritan	6008
12	marimacho	marimacho	5770
13	para ser mujer	to be a woman	4693
14	nenaza	nenaza	4358
15	odio a las mujeres	I hate women	2749
16	lagartona	gold digger	2006
17	a las mujeres hay que	to the women we have to	1845
18	las mujeres no deberian	women shouldn't	1285
19	las mujeres de hoy en dia	women these days	991
20	mujer al volante	woman driving	962
21	mucho feminismo pero	a lot of feminism but	852
22	mujer tenias que ser	you had to be a woman	683
23	pareces una puta	you look like a bitch	474
24	para ser chica	to be a woman	180
25	acabaras sola	you'll end up alone	50
26	hombre que te aguante	man who endures you	37
27	obsesionada con el machismo	obsessed with sexism	8
28	pareces una fulana	you look like a whore	5
29	no ha probado un hombre	hasn't tried a man	1

Starting from these expressions, we used the Twitter API to search for tweets containing all selected keywords. Data was collected between July and December 2018, gathering 181792 tweets for terms listed in Table 1. The initial setup of our crawler implies collecting 100 tweets for each term daily, thus, ideally we would collect 2900 tweets per day. Using this methodology for corpus construction, we ensure that we obtain both sexist and non-sexist tweets for each keyword expression. For example, even the expression "a fregar" ("*go washing*") is commonly used to under-valuate women's capacity to work, it also occurs in non-sexist tweets and our dataset is expected to reflect this ambiguity.

We established two constraints to the data collection process. On the one hand, we limited the collection of tweets to 15000 per keyword expression. On the other hand, we set a minimum threshold of 150 tweets per expression. After collecting the information, we took a random sample of 150 tweets per term. We verified that tweets had a difference of at least one day to avoid conversations and to ensure data is spread over the six months. Table 1 shows the number of tweets collected per term. We discarded 5 terms since they do not reach 150 tweets. As mentioned above, we sample the original dataset to build the final corpus composed by 3600 tweets.

[5]https://github.com/franciscorodriguez92/MeTwo
[6]Ana Isabel Bernal-Triviño Twitter account: @anaisbernal

TABLE 2. Terms selected to build MeTwo together with examples of sexist and non-sexist tweets extracted from the crawled tweets.

Group	Terms	Category	Example of sexist tweet	Example of non-sexist tweet
1	feminazi, las feministas, mucho feminismo pero, obsesionada con el machismo	Ideological discredit	Las feministas de esta época forman círculos cerrados, levantan banderas caprichosas y difíciles de ondear, se defienden entre sí, no el género de manera conjunta, y no ven más allá de su núcleo o colectivo	¿No te convencen mis argumentos? Intentemos debatir. ¿Usas "feminazi"? Te quedas solo
2	a la cocina, a fregar	Role stereotyping	@user Nos pone a fregar rápido, como tiene que hacer una mujer de su casa digna de su hijo.	Me dio sed, pero me da miedo ir a la cocina
3	como una mujer, como una niña, nenaza, para ser mujer, para ser chica, mujer tenías que ser, mujer al volante	Inferiority/incapacity	Mujer al volante, tenga cuidado!	@user El dia que esta persona pueda tener un bebe vaginalmente, amamantar, sentir y amar como una mujer sera una mujer
4	zorra, lagartona no ha probado un hombre, mojigata	Sexualization	Y CON ESA CARITA ME MOJIGATA QUE TIENE...:)...SON LAS PEORES...	Alguien me explica que zorra hace la gente en el cajero que se demora tanto.
5	A las mujeres hay que, las mujeres no deberían, las mujeres de hoy en día	Male dominance	Por eso es que las mujeres no deberían tener derecho al voto	Las mujeres no deberían maquillarse, son lindas con el simple hecho de existir.
6	loca del, niñata, odio a las mujeres, hombre que te aguante, en tus días	Hate and violence	dios mio como odio a las mujeres, una mas embrollera que la otra	No tengo ni 30 y ya soy la loca del té
7	marimacho, pareces una fulana, pareces una puta	Physical stereotyping	@user la dictadura de las feas, sobre las bonitas. Aún no conozco a una feminazi, mina o que no sea marimacho. Porque será?	m han llamado marimacho toda mi vida, pero eh ahora como mola tu rollo tia xd

TABLE 3. Terms selected to build MeTwo together with examples of sexist and non-sexist tweets extracted from the crawled tweets (English translation).

Group	Terms	Category	Example of sexist tweet	Example of non-sexist tweet
1	feminazi, the feminists, a lot of feminism but, obsessed with sexism	Ideological discredit	The feminists of this time form closed groups, raise flags that are difficult to wave, defend their organization but not the gender jointly and do not see beyond their collective.	Don't my arguments convince you? Let's try to debate. Do you use "feminazi"? You stay alone
2	to the kitchen, go washing	Role stereotyping	@user makes us wash immediately, as a woman in her house worthy of her child has to do.	I'm thirsty, but I'm scared to go to the kitchen
3	like a woman, like a girl, nenaza, to be a woman, to be a girl, you had to be a woman, woman driving	Inferiority/incapacity	Woman driving, be careful!	@user The day this person can have a baby vaginally, breastfeed, feel and love like a woman will be a woman
4	slut, gold digger, hasn't tried a man, puritan	Sexualization	AND WITH THAT PURITAN FACE THAT SHE HAS...:)...THEY ARE THE WORST...	Someone explains to me what the fuck makes the people at the cashier take so long.
5	To the women we have to, women shouldn't, women today	Male dominance	That's why women should not vote	Women should not put on makeup, they are cute with the simple fact of existing.
6	crazy about, little girl I hate women, man who endures you you'll end up alone in your days	Hate and violence	My God, how I hate women, one more confusing than the other	I'm not even 30 and I'm already crazy about tea
7	marimacho, you look like a bitch, you look like a whore	Physical stereotyping	the dictatorship of the ugly, over the beautiful. I still don't know a feminazi who isn't a marimacho. Why is that?	I've been called marimacho all my life, but now: hey your style is so cool lol

After this, we manually examined the dataset in order to understand the different ways that sexism is expressed and the different facets of women that are most frequently undermined or criticised in online communications via social networks such as Twitter. Table 2 (which is translated to English in Table 3) shows the sexist terms and expressions grouped by their category and semantic meaning, along with examples for both sexist and non-sexist tweets that include such terms. It is important to note that this grouping of tweets is not meant to be an exhaustive categorization of sexist expressions and is only based on the empirical observation of the dataset.

Group 1 gathers terms and expressions related to **ideological sexism**. Terms in this group try to underestimate feminism and the struggle of women for equality. For instance, the term ''feminazi'' is widely used on social media to attach

negative connotations to feminism by comparing it to nazism. An example of "feminazi" in a sexist context can be shown in the tweet "*Uy habló de inventos la feminazi*" (*"Oops talking about lies the feminazi"*). Similarly, the tweet "*Las feministas de esta época forman círculos cerrados, levantan banderas caprichosas y difíciles de ondear, se defienden entre sí, no el género de manera conjunta, y no ven más allá de su núcleo o colectivo.*" (*"The feminists of this time form closed groups, raise flags that are difficult to wave, defend their organization but not the gender jointly and do not see beyond their collective."*) from table 2 is underestimating feminism. Because of the problem of word ambiguity, we can also find this term in non-sexist contexts like in the sentence "*?'No te convencen mis argumentos? Intentemos debatir. ?'Usas "feminazi"? Te quedas solo*" ("*Don't my arguments convince you? Let's try to debate. Do you use "feminazi"? You stay alone*").

Group 2 collects terms associated with **role stereotyping**, in particular, those suggesting that women are meant to do the housework and parenting. The tweet "*@user Nos pone a fregar rápido, como tiene que hacer una mujer de su casa digna de su hijo.*" ("*@user makes us wash immediately, as a woman in her house worthy of her child*") is an example of sexist expression in this group.

Terms in **group 3** underestimate the ability of women to carry out some tasks (**intellectual incapacity and inferiority**). The sentence "*Mujer al volante, tenga cuidado!*" (*Woman driving, be careful!*) underestimates women's ability to drive.

Expressions in **group 4** aim to **sexualize and objectify women**. This group collects terms that suggest women have sex in exchange for money or favors, or that consider that women must be sexually active and willing to satisfy the sexual needs of men. The sentences *"No le presten atencion. como yo hago a cualquier otra zorra"* (*Do not pay attention. as I do to any other slut*) and *"Y CON ESA CARITA DE MOJIGATA QUE TIENE…:)…SON LAS PEORES…"* (*AND WITH THAT PURITAN FACE THAT SHE HAS…:)…THEY ARE THE WORST…*) use terms *"zorra"* (*slut* or *bitch*) and *"mojigata"* (*puritan*) in a sexist context.

Group 5 captures terms expressing **patriarchy behaviours and male dominance**. These terms are used to suggest that men and women play different roles in society and that men deserve greater privileges than women. For instance, the sentence *"Por eso es que las mujeres no deberían tener derecho al voto"* ("*That's why women should not vote*") suggests that only men should be allowed to decide at the polls.

Group 6 gathers terms that are frequently used to express **hate and violence against women**. It includes tweets that explicitly express hatred towards them. It is also important to clarify that the term **"niñata"** (*little girl*) is used to indicate that a woman is immature so it does not necessarily have sexist connotations. Figure 1 shows how most of the tweets containing this term are annotated as non-sexist. The sentence "*dios mio como odio a las mujeres, una mas embrollera que la otra*" (*My God, how I hate women, one more confusing than the other*) would be an example of this category. According to the definition of misogyny as hatred or prejudice against women, this group would be the closest to it.

Finally, **group 7** contains terms used to suggest that women should take care of their **physical appearance**. It is important to say that "marimacho" can be also used as a homophobic term since it suggests a woman has masculine features but it is not sexist *per se*. However, it can be also used in a sexist context like in sentence *"@user la dictadura de las feas, sobre las bonitas. Aún no conozco a una feminazi, mina o que no sea marimacho. Porque será?"* (*the dictatorship of the ugly, over the beautiful. I still don't know a feminazi who isn't a marimacho. Why is that?*). On the other hand, the sentence *"me han llamado marimacho toda mi vida, pero eh ahora como mola tu rollo tia xd"* (*I've been called marimacho all my life, but now: hey your style is so cool lol*) represents a non-sexist example.

B. ANNOTATION AND AGREEMENT

We propose the following labels to identify sexist expressions and behaviours in Twitter:

- **SEXIST**: tweets that underestimate women as a result of their gender, independently of the facet of women that is criticised, and independently of the intentionality and violence. Example: "*@user Lo irónico es que lo dice una mujer, que naturalmente debería callarse y dedicarse a la cocina, limpiar y criar hijos*" ("*@user The irony is that it is being said by a woman, which naturally should shut up and devote herself to cooking, cleaning and raising children*").
- **NON-SEXIST**: tweets without sexist connotations. In this category, we could find xenophobic or offensive tweets but that do not understimate women for the reason of their gender. For instance: "*@user @user POR CIERTO, EN TU FOTO DE PERFIL SE PUEDE OBSERVAR QUE ERES BASTANTE VARONIL, ASÍ QUE SI NO ERES MARIMACHO, EMPIEZA A SERLO*" ("*by the way, in your profile picture you can see that you are quite manly, so if you are not a marimacho, yo should start to be*").
- **DOUBTFUL**: tweets that could be sexist depending on the context, which can not be inferred from the text in the tweet. In this category we find tweets that would be sexist if they were specifically targeted to women, such as, for example: "*@user @user Más vale que se marche a fregar!*" ("*@user @user You better go washing!*").

MeTwo was labeled based on a majority vote by three annotators. In case of total disagreement, a fourth annotator decided the final label. Of the 3600 tweets, the fourth annotator was only required in 20 tweets. Given the subjectivity and difficulty of the task, we developed an annotation guide in which we provided a clear explanation of each label along with a number of examples.

During the annotation process, some difficulties were found due to language phenomena such as irony or

TABLE 4. Agreement between annotators (average of kappa coefficient).

	Kappa
Annotator 1-2	0,68
Annotator 1-3	0,68
Annotator 2-3	0,88
Average	**0,75**

TABLE 5. Total agreement by label.

Label	Agreement (%)
SEXIST	74 %
NON-SEXIST	87.2 %
DOUBTFUL	59.24 %

TABLE 6. Distribution of labels in the dataset.

Label	Tweets
NON-SEXIST	2181 (60.58%)
SEXIST	1152 (32%)
DOUBTFUL	267 (7.42%)

ambiguity. For instance, the sentence "*Mucho feminismo pero a la primera de cambio.…*"("*Much feminism but at the first opportunity.…..*") could be using irony making harder to detect a sexist attitude. A good example of the importance of ambiguity can be shown in the sentence "*@user La zorra guardando las gallinas. ¡ Que se encargue Rosell ¡¡ Bueno…, cuando salga de la cárcel. Cinismo en grado máximo.*" ("*@user The [fox/slut] keeping the chickens. ¡¡ Let Rosell take care Well … when he leaves the jail. Cynicism in maximum degree.*"). The word *"zorra"* in Spanish could be refered to *"slut"* or to *"fox"* so it is unclear if it is employed in a sexist context.

Another important problem found is related to tweets which cite sexist content [56]. One could argue than the tweets are not sexist *per se*, but they tell about sexist behaviours. A good example of this problem can be shown in the sentence "*Pareces una puta con ese pantalón. -Mi hermano de 13 cuando me vió con un pantalón de cuero*"("*You look like a whore with those pants. -My brother of 13 when he saw me wearing leather pants"*). In this work, all tweets containing sexist cites have been considered as sexist.

To assess the reliability of the annotation process, we evaluate the inter-annotator agreement. To measure this, we opted for the Cohen's kappa coefficient [55]. A poor value on this measure could indicate some problem in the annotation process. On the one hand, it may suggest that the task is too difficult or subjective, even for humans. On the other hand, it may indicate that annotators are not prepared enough to carry out the task. Final inter-annotator agreement results are shown in Table 4. We achieved a value of 0.75 for kappa coefficient. Kappa values greater than 0.6 are usually consider as substantial agreement and adequate for this type of task [57]. Moreover, in Table 5, the percentage of agreement for the three annotators by label is depicted. The "DOUBTFUL" label seems to be the most difficult to detect even for humans. This may be due to the fact that, in some cases, the existence of sexist context is subjective. For instance, the sentence *"Profesiones preferidas de las feministas: Psicóloga, crítica cultural, profesora universitaria, diputada."*("Preferred professions of feminists: Psychologist, cultural critic, university professor, congresswoman.") has been considered "DOUBTFUL" by two annotators and "SEXIST" by the other one. On the contrary, annotators seem mostly to agree on the "NON-SEXIST" and "SEXIST" labels.

The distribution of labels in the dataset is showed in Table 6. As it can be observed, there is an important bias towards the "NON-SEXIST" label, due to the nature of the problem.

FIGURE 1. Terms distribution by category.

Finally, Figure 1 shows the distribution of labels by term. Note that, as previously mentioned, "niñata" (*"little girl"*) and "marimacho" are considered sexist only in very specific contexts, so that tweets containing such terms are usually labeled as "NON-SEXIST". In contrast, we can see that terms such as "feminazi" or "nenaza" are mostly used in sexist contexts. Besides, there are terms, such as "zorra" (*slut* or *bitch*) o "a fregar" ("*go washing*"), for which the percentage of "DOUBTFUL" tweets is very high, given that these are polysemous words whose meaning strongly depends on the context.

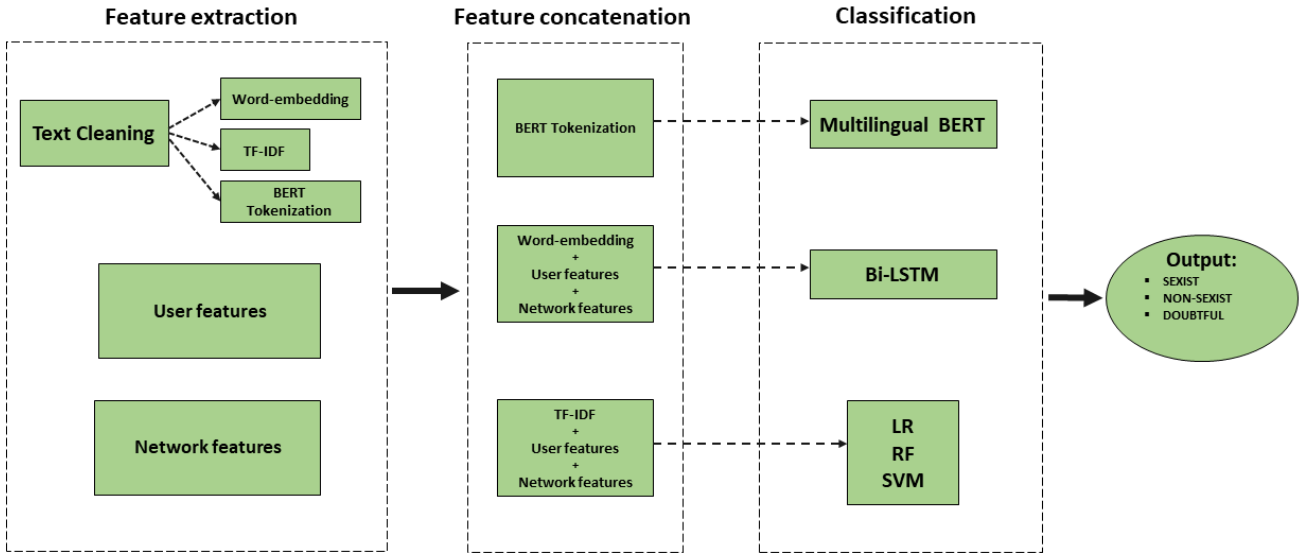


FIGURE 2. Classification system architecture.

IV. AUTOMATIC DETECTION OF SEXISM IN TWITTER

We propose to identify sexist speech on social media using a supervised classification system. The overall experiment plan which describes the three main approaches to solve the classification task is depicted in Figure 2.

In the first step, a preprocessing is carried out so that the text is cleaned and all relevant linguistic, user and network-based features are collected. To this end, different transformations to features are applied to feed the ML algorithms used. Next, features are combined differently depending on the classification model. Finally, the classification module uses the information gathered in the previous steps to identify sexist signals in text. We compare both traditional methods using tf-idf features and deep learning-based methods using word embeddings. The following subsections describe the classification module in detail. The code for all methods has been made available on Github.[7]

A. FEATURE EXTRACTION

In order to gather useful information for the classification algorithm, a preprocessing step is accomplished. Besides textual attributes, we experiment with other additional features in order to enrich the information available in the classification step. More specifically, features considered can be grouped into user-, network-, or text-based.

1) USER AND NETWORK-BASED FEATURES

Regarding user-based features, we use the following data that is directly extracted from the user's Twitter profile: followers count, friends count, number of lists registered, number of tweets posted, number of favorites registered, device used to post and presence of verification. Besides, we employ some features related to the network and how it interacts with the tweet. In particular, we use favourite count, retweet count, presence of hashtags, presence of URLs, presence of images or videos and presence of mentions.

2) TEXT-BASED FEATURES

Features extracted from the text of the tweets are, a priori, the most relevant for our classification system. In this group of features, we use two different ones: the text and the length of the tweet. Before extracting the textual signals, we apply the following preproceessing to the texts:

- Replacing emojis by a description.
- Replacing URLs by the keyword "twurl".
- Replacing user mentions by the keyword "twuser".
- Removing Spanish accents.
- Removing punctuation marks.
- Converting hashtags containing capital letters. For instance, we convert "#HappyBirthday" to "happy birthday".
- Replacing the rest of the hashtags by the keyword "twhashtag".
- Converting all letters to lowercase.
- Replacing exclamation marks by the keyword "twexclamation".
- Replacing question marks by the keyword "twinterrogation".
- Tokenizing the text of the tweet, using the NLTK api.[8]
- Removing stop words
- Normalizing and replacing slang, using a Spanish lexicon developed in [58].
- Stemming, using the Porter Stemmer [59].

Once all the preprocess has been applied to the text, the next step aims to build the features that will feed the

[7] https://github.com/franciscorodriguez92/code-sexism-detection-spanish

[8] https://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.casual

classifiers. We divide the text features in two categories: lexical (tf-idf features) and semantic (word embeddings).

On the one hand, traditional classification systems use tf-idf vectors, which are built from the tweets unigrams. We explore the use of bigrams and trigrams without any improvement. Tf-idf features are created using terms whose document frequency is, at least, 1%. On the other hand, deep learning methods use a word embedding approach to detect semantic and syntactic words relations. In this paper, all the pre-trained embeddings are built using the word2vec algorithm [60]. This method allows us to assign a static numeric vector to each token.

We use three different pre-trained word embedding datasets. Firstly, we employ an embedding resource trained on the Spanish Billion Word Corpus (SBWCE) [61]. This dataset consists of more than one million word embeddings of dimension 300. Secondly, we use a pre-trained resource trained on tweets [62] using a vector length of 200 dimensions (WESB). Finally, we use a pre-trained word embedding resource trained on the Spanish CoNLL17 corpus [63] composed of 300-dimensional vectors (CoNLL17).

B. CLASSICAL CLASSIFICATION METHODS

A number of classical machine learning techniques can be used for this task. We opted for Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) since they are widely used for this type of task [14], [50], [64], [65].

As far as the feature extraction is concerned, user and network-based features are used along with TF-IDF attributes. For every classifier, default hyper-parameters are employed so that they are not tuned using a sample of MeTwo. We have explored hyper-parameters tuning using a sample of data without success.

C. BIDIRECTIONAL LONG SHORT-TERM MEMORY (BI-LSTM)

Approaches based on neural networks have been successfully used for Natural Language Processing tasks [28], [54]. In this paper, we compare some deep neural network approaches to the classical ones. In particular, we experiment with deep recurrent neural networks such as Bidirectional Long Short-Term Memory (Bi-LSTM) [66]. We use Bi-LSTMs to capture long range dependencies in tweets, which may not be captured by classical classification algorithms.

Our Bi-LSTM model architecture is depicted in Figure 3. The first layer of the network performs word embedding. We experiment four different configurations for this layer. First, we experiment with an embedding layer that is trained along with all the network. For all other experiments, we use the pre-trained embeddings described in subsection IV-A2.

The next layer is composed by a Bi-LSTM module with 200 units (LSTM cells) as input and 50-dimensional vectors as output. Then, we apply max pooling operation since we get all hidden states in previous Bi-LSTM layer. Next, a dropout layer (0.1 dropout rate) is applied

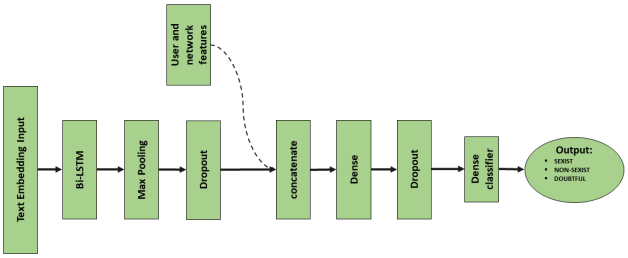


FIGURE 3. Bi-LSTM architecture.

to avoid over-fitting. After that, we concatenate user and network-based features to the output of previous layer and add a fully-connected layer (dense layer).

Finally, we add a dropout layer (0.1 dropout rate) and a fully-connected output layer with one neuron per predicted class (three neurons in total), and a sigmoid activation to normalize output values. The Adam optimizer [67] is used along with binary cross entropy as loss function.

D. BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

In the previous section, one of the most frequently used context-independent neural embedding is applied. Here, we apply a context-dependent and transformer-based language model called BERT [31]. In particular, we use BERT-Base Multilingual Cased[9] which provides sentence representations for 104 languages. Since its publication, many positive experimental results have been published in different tasks [68], [69]. To perform the sexism detection task, we fine-tune the pre-trained mBERT-Base parameters adding a fully-connected layer on top of mBERT to minimize loss function in our particular task. In our experiments, we trained our classifier with a batch size of 16 for 25 epochs. The dropout probability is set to 0.1 for all layers. The Adam optimizer is used with a learning rate of 2^{-5} and a weight decay of 0.01. As an input, we tokenized each tweet using the BERT tokenizer.

V. RESULTS AND ANALYSIS

In this section, we analyze and discuss the results obtained in all different experiments. We also include the details of our evaluation methods and error analysis.

A. EXPERIMENTAL SETUP

For our experiments, we use classical and neural network machine learning libraries in Python. For the implementation of Bi-LSTM, Keras[10] was used. On the other hand, we used pytorch [70] for the implementation of mBERT. In particular, the transformers library from huggingface was employed [71]. Finally, the scikit-learn[11] library was used to implement traditional classification systems.

[9]https://github.com/google-research/bert/blob/master/multilingual.md
[10]https://keras.io/
[11]https://scikit-learn.org/

TABLE 7. Results for sexism detection for proposed methods.

	Accuracy	F1	Recall	Precision
Baseline	0.61	0.2	0.3	0.24
Baseline (tf-idf)	0.68	0.59	0.62	0.59
LR	0.71	0.62	0.64	0.62
SVM	0.71	0.61	0.63	0.61
RF	0.72	0.6	0.57	**0.67**
Bi-LSTM	0.71	0.61	0.62	0.6
Bi-LSTM + SBWCE	0.72	0.62	0.65	0.62
Bi-LSTM + WESB	**0.74**	**0.64**	**0.66**	0.63
Bi-LSTM + CoNLL17	0.73	**0.64**	0.65	0.64
mBERT	0.73	0.63	**0.66**	0.62
mBERT (text features)	**0.74**	**0.64**	**0.66**	0.63

B. SEXISM DETECTION RESULTS

Here, we report the performance of our proposed approaches in comparison with the baselines. Two baselines have been proposed: the first one is based on tf-idf features along with the LR classifier; and the second one labels each record based on the majority class.

The performance evaluation of all classifiers followed a 10-fold cross-validation. The effectiveness is measured using four different metrics: accuracy, precision, recall and F1-score. Table 7 summarized the results obtained for each classification method.

According to Table 7, mBERT with text features and Bi-LSTM with WESB embeddings seem to perform the best. Intuitively, this makes sense since BERT yields excellent results in many tasks related to NLP [31] and is able to gather information containing syntactical and contextual features. On the other hand, pre-trained embeddings from WESB were learned using tweets, which could be more appropriate to represent tweets from MeTwo.

All methods outperform our baselines in all metrics. As show in Table 7, all classical learning models achieve comparable results with the exception of random forest, whose high value in precision is due to the bias towards "NON-SEXIST" tweets in the class distribution (see Table 6). Traditional methods also perform similarly to Bi-LSTM without pre-trained embeddings. By contrast, combining Bi-LSTM with pre-trained word embeddings improves the performance of the method up to 3% in F1-score. Similarly, all versions of mBERT outperform traditional methods.

Regarding neural networks methods, mBERT and Bi-LSTM with pre-trained word embeddings achieve comparable results. It is important to state that adding user and network-based features does not lead to any improvement for mBERT. This could be due to the fact that we are just concatenating these extra features and adding a simple linear layer on top of BERT to fine-tune the model [72]. A more exhaustive study should be done to determine the influence of extra features and how to combine them effectively when using BERT.

C. ERROR ANALYSIS

Although we achieve interesting results and improvements with respect to the baselines, all models are still making some mistakes. To understand better the source of the failures, we have performed a deep analysis on model errors. In particular, we further investigate results of the mBERT model.

TABLE 8. Confusion matrix for the mBERT classifier.

True/Predicted	NON SEXIST	SEXIST	DOUBTFUL
NON SEXIST	1808	321	52
SEXIST	360	745	47
DOUBTFUL	89	68	110

TABLE 9. Accuracy by group and term.

Group	Accuracy	Term	Accuracy
1	0.62	feminazi	0.78
		las feministas (*the feminist*)	0.51
		mucho feminismo pero (*a lot of feminism but*)	0.57
2	0.75	a la cocina (*go to the kitchen*)	0.85
		a fregar (*go washing*)	0.66
3	0.8	como una mujer (*as a woman*)	0.83
		como una niña (*as a girl*)	0.93
		nenaza	0.94
		para ser mujer (*to be a woman*)	0.87
		para ser chica (*to be a woman*)	0.71
		mujer tenías que ser (*you had to be a woman*)	0.64
		mujer al volante (*woman driving*)	0.74
4	0.6	lagartona (*gold digger*)	0.43
		mojigata (*puritan*)	0.89
		zorra (*bitch*)	0.48
5	0.65	a las mujeres hay que (*to the women we have to*)	0.63
		las mujeres no deberían (*women shouldn't*)	0.65
		las mujeres de hoy en día (*women these days*)	0.68
6	0.83	en tus dias (*in your days*)	0.81
		loca del (*crazy about*)	0.87
		odio a las mujeres (*I hate women*)	0.71
		niñata (*little girl*)	0.95
7	0.81	pareces una puta (*you look like a bitch*)	0.68
		marimacho	0.95

Table 8 shows the confusion matrix for mBERT when only textual features are used. Note that "NON-SEXIST" tweets are easier to classify than "DOUBTFUL" and "SEXIST". "NON-SEXIST" tweets are correctly classified with 85% of accuracy. For "DOUBTFUL" class, almost 59% of tweets are misclassified. Intuitively, this makes sense since "DOUBTFUL" is the minority class and our model does not have many instances of this type during fine-tuning process. However, given that our ultimate goal is to detect sexism, the mistakes made for "SEXIST" class are critical. For this category, a 65% of accuracy is achieved.

We have observed that the performance could depend on the number of instances available for each class. Therefore, it seems that increasing the number of instances for "SEXIST" class would improve its detection. It is worth remembering that our work does not make use of any external resource of sexist vocabulary. As mentioned, previous works have demonstrated the benefit of using sexist lexicon. Detecting sexist words and including them in the model could help to reduce the misclassification of "SEXIST" tweets and will be investigated in future work.

To further investigate errors when detecting "SEXIST" tweets, we analyze the performance for each seed term. Table 9 shows accuracy by group of sexist terms and by term. Groups 6 and 7 achieve the highest accuracy. On the one hand, group 6 is composed by terms expressing hate and

TABLE 10. Examples labeled by the model.

N	text	Real	Predicted
1	@user Tu eres Tonta, ponte a fregar anda (*@user You are dumb, go washing*)	SEXIST	NON SEXIST
2	Nunca voy a entender a las mujeres de hoy en día @url (I will never understand women these days @url)	SEXIST	NON SEXIST
3	-Es que yo no sirvo para ser madrasta- :) las mujeres de hoy en día! (*-I do not serve to be a stepmother- :) women these days*)	SEXIST	NON SEXIST
4	A LAS MUJERES HAY QUE TRATARLAS BIEN Y BONITO, POR QUE SI NO, SE ENAMORAN. (*WOMEN MUST BE TREATED WELL AND PRETTY, BECAUSE IF NOT, THEY FALL IN LOVE*)	SEXIST	NON SEXIST
5	Tooooodas nomas sufrieron violencia se generó que esto lo otro , que onda las mujeres de hoy en dia *(All suffered gender violence, women these days)*	SEXIST	NON SEXIST
6	@user Zorra jajaaja te amo (*@user bitch lol I love you*)	NON SEXIST	SEXIST
7	Alguien me explica que zorra hace la gente en el cajero que se demora tanto. (*Someone explains to me what the fuck people do at the ATM that takes so long.*)	NON SEXIST	SEXIST

misogyny thus they were expected to be easier to detect, since this type of hateful messages make use of a violent vocabulary that is not usually employed in ''NON-SEXIST'' tweets. On the other hand, group 7 achieves good performance because of the unbalanced distribution of terms which it gathers (Figure 1). In contrast, group 5, which represents the expression of male dominance, presents the second worst performance. This may be due to the fact that sexism in this group is more subtle and the sexist ideas are expressed implicitly, using expressions and words that are highly dependent on the context. The poor performance in group 4 is due to the balanced distributions in its terms. The existence of many ''DOUBTFUL'' tweets makes more difficult the task since the classifier performs poorly when detecting this class.

Regarding the accuracy by term, ''niñata'' and ''marimacho'' achieve the highest accuracy. This is because such terms are highly biased to the ''NON-SEXIST'' label (Figure 1). The amount of information should be increased so that we have more diversity of labels for such terms. In contrast, the expression ''lagartona'' has the worst performance. Most tweets containing this term express sexist behaviours which are not radical or explicit. Some of them gather subtle sexism which is more difficult to detect.

After this quantitative error analysis, a manual inspection has been carried out to better understand the problems of our classification system. Table 10 gathers some of the examples examined. In tweets 1 and 2 our classifier is not able to detect sexism, the reasons seem to be the tweet length and the lack of explicit sexist slung. There is a subset of tweets in which our classifier tends to fail because the short length of the text and the absence of terms highly used in sexist contexts. In fact, we have observed that misclassified tweets were, on average, 5 characters shorter than the average in the MeTwo dataset.

Again, some cases containing implicit sexism such as tweets 3, 4 and 5 are not detected by our classifier, the reasons being the use of irony and the lack of explicit sexist vocabulary [73]. Regarding tweet 4, note that the expression ''a las mujeres hay'' is biased towards the ''NON-SEXIST'' label (Figure 1). Moreover, the tweet contains the term ''*bonito*'' (''beautiful'') which is not typically employed in sexist contexts. It is also important to note that our strategy to construct the MeTwo dataset is keyword-based, which can introduce natural biases towards certain sexist terms. Bias mitigation techniques will be investigated in future works [74].

TABLE 11. Generalization experiment.

	Training/evaluation	Class	F1	Recall	Precision
SVM	MeTwo/AMI	Sexist	0.64	0.53	0.81
		Non-sexist	0.36	0.58	0.26
		Macro Avg.	**0.55**	**0.55**	**0.54**
	AMI/MeTwo	Sexist	0.5	0.67	0.41
		Non-sexist	0.45	0.35	0.61
		Macro Avg.	0.48	0.51	0.51
mBERT	MeTwo/AMI	Sexist	0.49	0.6	0.42
		Non-sexist	0.63	0.56	0.73
		Macro Avg.	**0.56**	**0.57**	**0.58**
	AMI/MeTwo	Sexist	0.46	0.38	0.59
		Non-sexist	0.57	0.69	0.48
		Macro Avg.	0.51	0.53	0.53

D. SEXISM DETECTION VS. AUTOMATIC MISOGYNY DETECTION (AMI)

At this point, we wonder how well a model trained on our corpus generalizes to other datasets from a subdomain of sexism, such as misogyny. We have hypothesized that sexism includes misogyny, hence MeTwo should be able to capture some misogynistic attitudes and behaviours. To estimate this, we train two models on MeTwo and evaluate in the AMI dataset used for the IberEval competition, and vice versa.

The AMI dataset used for this experiment (training set) is composed by 3307 tweets and considers only two classes: ''misogynous'' and ''not misogynous''. Therefore, to make the comparison, we remove all tweets from ''DOUBTFUL'' class in MeTwo and use the ''SEXIST'' and ''NON-SEXIST'' label to perform this experiment (3333 tweets). We performed two different experiments with SVM and mBERT in order to compare traditional with neural networks based methods.

Table 11 shows the results of this experiment. We report precision, recall and F1 for every class, and also a macro average for all classes. We do not report accuracy since both datasets are not fully balanced and would not be useful for comparison. It can be observed that, performance training on MeTwo is higher in all macro average metrics, hence evidencing that a model trained on MeTwo is able to generalize better to AMI than vice versa. More noticeably, for the ''SEXIST'' class, MeTwo outperforms AMI for both classifiers.

These results show that MeTwo gathers sexist attitudes in a broad sense, from misogyny to other types of sexism. Thus, since MeTwo contains misogyny, the model trained on it is able to correctly classify some of the information present in AMI. In contrast, since AMI just contains tweets expressing misogyny or hatred towards women, the model works worse when evaluating on MeTwo, a dataset of broad sexism.

Additionally, we perform a last experiment using the MeTwo corpus as training and only the test set of AMI for evaluating the performance, so that we can compare our results to the ones obtained during the AMI competition.

TABLE 12. Generalization experiment on AMI test set.

	Training	Acc.	F1	Recall	Precision
SVM	MeTwo	0.58	0.57	0.58	0.58
mBERT	MeTwo	**0.65**	0.65	0.65	0.65
First team	AMI	**0.81**	-	-	-
baseline	AMI	0.77	-	-	-
Last team	AMI	0.53	-	-	-

Table 12 shows the results of this experiment as well as the best and worst team, and baseline in the AMI competition. Since the task organizers only provide the accuracy score as the competition baseline, we will use this metric for comparison. In terms of accuracy, our best system would rank 23 out of 25 teams with an accuracy of 0.65 [49].

It must be noted that, since they share data collection approaches, there is a strong relationship between both test and training sets unigrams in AMI. Some of the most frequent unigrams for misogynistic tweets in both sets are "*puta*" (bitch), "*perra*" (bitch/slut) or "*callate*" (shut up). However, the most frequent unigrams for sexist tweets in MeTwo are "*mujer*" (woman), "*mujeres*" (women), "*feminismo*" (feminism). Therefore, it was expected that a classifier trained on the training set of AMI will outperform the one trained on the MeTwo training set when evaluating on the test set of AMI. This is due to the fact that, even if both datasets are topic-related, the performance of this experiment is heavily influenced by the data collection approaches. In addition to this, we observe that the best results for AMI competition were achieved using hate external lexicons. It suggests that the use of hate slurs (less important in MeTwo) is a clear signal of misogynistic content.

Nonetheless, even if we use for training a dataset that was collected differently and for a different purpose than the one used for testing, we still rank better than other participants to the AMI challenge. It indicates that MeTwo is not just composed of subtle and non-hateful tweets, but it also gathers some misogynistic messages and can predict some of the AMI tweets correctly.

VI. CONCLUSION AND FUTURE WORK

In this paper, we perform an exhaustive analysis to understand how sexist attitudes and behaviours are expressed in social networks conversations. In particular, our aim is to detect sexism in a broad sense in Twitter. To this end, we presented an automatic system based on ML that allows us to compare traditional and neural networks based methods. Regarding feature extraction, we compared methods based on traditional tf-idf features to word embeddings approaches. The experimental results showed that BERT outperforms the rest of algorithms tested achieving an accuracy of 74% in the detection of sexist expressions. On the other hand, MeTwo was presented, the first Spanish corpus of sexism expressions in Twitter. The corpus is intended to cover a broad spectrum of sexist attitudes, from subtle and non-hateful sexism to misogyny and radical sexism. Metwo comprises 3600 tweets labeled as sexist or not based on a majority vote by three annotators. To the best of our knowledge, this is the first resource of this type for Spanish texts. We have also made a preliminary attempt to group sexist expressions in categories according to the different facets of women that are attacked. Furthermore, an error analysis has been performed to understand the limitations of our system, that are mainly due to linguistic phenomena such as irony and word ambiguity as well as to the lack of enough context and the size of the dataset.

As future work, we plan to extend our system to detect different types of sexism such as sexualization or role stereotyping (see Tables 2 and 3). To this aim, a new exhaustive categorization of sexist expressions should be done. This may help to examine how the different types of sexism are expressed and how they spread through social networks. Besides, we will incorporate a multi-lingual approach to handle different languages at the same time. To this end, a new corpus will be created to include more data sources and languages. Finally, new signals, such as the use of sexism slurs, will be explored.

REFERENCES

[1] M. F. Wright, B. D. Harper, and S. Wachs, "The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition," *J. Personality Individual Differences*, vol. 140, pp. 41–45, Apr. 2019.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. ICWSM*, 2017, pp. 1–4.

[3] D. Gershgorn and M. Murphy. (2017). *Facebook is Hiring More People to Moderate Content than Twitter has at Its Entire Company Quartz*. Accessed: Jun. 20, 2019. [Online]. Available: https://bit.ly/2ZbhsHu

[4] T. Vega, "Facebook says it failed to bar posts with hate speech," The New York Times, 2013. Accessed: Jun. 10, 2019. [Online]. Available: https://nyti.ms/2VXy9Ex

[5] R. Meyer, "Twitter's famous racist problem," The Atlantic, 2016. Accessed: Jul. 5, 2019. [Online]. Available: https://bit.ly/38EnFPw

[6] M. Duggan, "Online harassment 2017," Internet Tech., Pew Res. Center, Washington, DC, USA, Tech. Rep., Jul. 2017.

[7] R. Fulper, G. L. Ciampaglia, E. Ferrara, Y. Ahn, A. Flammini, F. Menczer, B. Lewis, and K. Rowe, "Misogynistic language on Twitter and sexual violence," in *Proc. ChASM*, 2015, pp. 1–4.

[8] A. Dhrodia, "Social media and the silencing effect: Why misogyny online is a human rights issue," NewStatesman, 2017. Accessed: Sep. 25, 2020. [Online]. Available: https://bit.ly/3n3ox68

[9] A. Karami, C. N. White, K. Ford, S. Swan, and M. Yildiz Spinel, "Unwanted advances in higher education:Uncovering sexual harassment experiences in academia with text mining," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102167.

[10] D. E. J. Austin and M. Jackson, "Benevolent and hostile sexism differentially predicted by facets of right-wing authoritarianism and social dominance orientation," *Personality Individual Differences*, vol. 139, pp. 34–38, Mar. 2019.

[11] M. Thelwall and E. Stuart, "She's reddit: A source of statistically significant gendered interest information?" *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1543–1558, Jul. 2019.

[12] T. Donoso-Vázquez, *Violencias de Génereo 2.0*, 1st ed. Barcelona, Spain: Kit-Book, 2014, pp. 13–27.

[13] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13825–13835, Feb. 2018.

[14] Z. Waseem, "Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter," in *Proc. 1st Workshop NLP Comput. Social Sci.*, 2016, pp. 138–142.

[15] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.

[16] Z. Ryan Shi, C. Wang, and F. Fang, "Artificial intelligence for social good: A survey," 2020, *arXiv:2001.01818*. [Online]. Available: http://arxiv.org/abs/2001.01818

[17] A. Khatua, E. Cambria, and A. Khatua, "Sounds of silence breakers: Exploring sexual violence on Twitter," in *Proc. ASONAM*, 2018, pp. 397–400.

[18] A. Khatua, E. Cambria, K. Ghosh, N. Chaki, and A. Khatua, "Tweeting in support of LGBT? A deep learning approach," in *Proc. ACM India Joint Int. Conf. Data Sci. Manage. Data*, 2019, pp. 342–345.

[19] E. Cambria, P. Chandra, and A. Hussain, "Do not feel the trolls," in *Proc. SDoW Workshop 9th Int. Semantic Web Conf.*, 2010, pp. 1–12.

[20] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Trans. Comput. Social Syst.*, early access, Sep. 17, 2020, doi: 10.1109/TCSS.2020.3021467.

[21] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2012, pp. 1980–1984.

[22] P. Gianfortoni, D. Adamson, and C. Rose, "Modeling of stylistic variation in social media with stretchy patterns," in *Proc. EMNLP*, 2011, pp. 49–59.

[23] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016.

[24] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in *Proc. ITASEC*, 2017, pp. 86–95.

[25] N. D. Gitari, Z. Zhang, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *Int. J. Multimedia Ubiquitous Eng.*, vol. 10, no. 4, pp. 215–230, Apr. 2015.

[26] Y. Tang and N. Dalzell, "Classifying hate speech using a two-layer model," *Statist. Public Policy*, vol. 6, no. 1, pp. 80–86, Jan. 2019, doi: 10.1080/2330443X.2019.1660285.

[27] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Detecting offensive language in tweets using deep learning," 2018, *arXiv:1801.04433*. [Online]. Available: http://arxiv.org/abs/1801.04433

[28] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. ACM WWW*, 2017, pp. 759–760.

[29] S. Zimmerman, U. Kruschwitz, and C. Fox, "Improving hate speech detection with deep learning ensembles," in *Proc. LREC*, 2018, pp. 1–8.

[30] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 41–45.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[32] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[33] 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.

[34] G. H. Paetzold, M. Zampieri, and S. Malmasi, "UTFPR at SemEval-2019 task 5: Hate speech identification with recurrent neural networks," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 519–523.

[35] M. Konstantinovsky. (2019). *What's the Difference Between Misogyny and Sexism howstuffworks?* Accessed: Jul. 13, 2020. [Online]. Available: https://people.howstuffworks.com/misogyny-and-sexism.htm

[36] K. Manne, *Down Girl: The Logic of Misogyny*. London, U.K.: Oxford Univ. Press, 2018.

[37] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," *J. Inf. Sci.*, May 2020, doi: 10.1177/0165551520917651.

[38] S. Sharifirad and A. Jacovi, "Learning and understanding different categories of sexism using convolutional neural network's filters," in *Proc. ACL*, 2019, pp. 21–23.

[39] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, and V. Varma, "Multi-label categorization of accounts of sexism using a neural framework," in *Proc. EMNLP-IJCNLP*, 2019, pp. 1642–1652.

[40] E. Fersini, F. Gasparini, S. Corchs, S. O. Textual, "Detecting sexist MEME on the Web: A study on textual and visual cues," in *Proc. ACIIW*, 2019, pp. 226–231.

[41] F. Gasparini, I. Erba, E. Fersini, and S. Corchs, "Multimodal classification of sexist advertisements," in *Proc. 15th Int. Joint Conf. e-Bus. Telecommun.*, 2018, pp. 565–572.

[42] J. Cardiff and E. Shushkevich, "Misogyny detection and classification in English tweets:the experience of the ITT team," in *Proc. EVALITA*, 2018, p. 182.

[43] D. Nozza, C. Volpetti, and E. Fersini, "Unintended bias in misogyny detection," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Oct. 2019, pp. 149–155.

[44] M. Anzovino, E. Fersini, and P. Rosso, "Automatic identification and classification of misogynistic language on Twitter," in *Proc. NLDB*, 2018, pp. 57–64.

[45] 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.

[46] M. ElSherief, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *Proc. AAAI*, 2018, pp. 1–10.

[47] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proc. EMNLP-IJCNLP*, 2019, pp. 4667–4676.

[48] O. El Ansari, Z. Jihad, and M. Hajar, "A dataset to support sexist content detection in arabic text," *J. Image Signal Process.*, vol. 12119, pp. 130–137, Sep. 2020.

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

[50] J. S. Canós, "Misogyny identification through SVM at IberEval 2018," in *Proc. IberEval*, 2018, pp. 229–233.

[51] V. Nina-Alcocer, "AMI at IberEval2018 automatic misogyny identification in Spanish and English tweets," in *Proc. IberEval*, 2018, pp. 274–279.

[52] S. Frenda and B. Ghanem, "Exploration of misogyny in Spanish and English tweets," in *Proc. IberEval*, 2018, pp. 260–267.

[53] E. W. Pamungkas, "Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets," in *Proc. IberEval*, 2018, pp. 234–241.

[54] I. Goenaga, A. Atutxa, K. Gojenola, A. Casillas, A. D. de Ilarraza, N. Ezeiza, M. Oronoz, A. Pérez, and O. Perez-de-Viñaspre "Automatic misogyny identification using neural networks," in *Proc. IberEval*, 2018, pp. 249–254.

[55] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[56] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully, "An annotated corpus for sexism detection in French tweets," in *Proc. ACL*, 2020, pp. 1397–1403.

[57] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[58] H. Gómez-Adorno, "Compilación de un lexicón de redes sociales para la identificación de perfiles de autor," *J. Res. Comput. Sci.*, vol. 115, pp. 19–27, May 2016.

[59] M. F. Porter, "An algorithm for suffix stripping," *J. Program.*, vol. 14, pp. 130–137, Jul. 1980.

[60] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS*, 2013, pp. 3111–3119.

[61] C. Cardellino. (Aug. 2019). *Spanish Billion Words Corpus and Embeddings*. Accessed: Dec. 10, 2019. [Online]. Available: https://crscardellino.github.io/SBWCE/

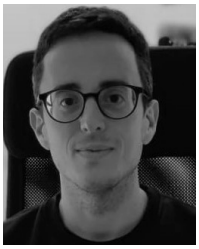
[62] J. Deriu. *Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification*. Accessed: Nov. 14, 2019. [Online]. Available: https://www.spinningbytes.com/resources/wordembeddings/

[63] M. Fares, M. Fares, S. Oepen, and E. Velldal, "Word vectors, reuse, and replicability: Towards a community repository of large-text resources," in *Proc. NoDaLiDa/WS*, 2017, pp. 271–276.

[64] H. Liu, F. Chiroma, and E. Haig, "Identification and classification of misogynous tweets using multi-classifier fusion," in *Proc. IberEval*, 2018, pp. 268–273.

[65] E. Shushkevich and J. Cardiff, "Classifying misogynistic tweets using a blended model: The AMI shared task in IBEREVAL 2018," in *Proc. IberEval*, 2018, pp. 255–259.

[66] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.
[67] P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[68] Y. Wang, W. Che, J. Guo, Y. Liu, and T. Liu, "Cross-lingual BERT transformation for zero-shot dependency parsing," in *Proc. EMNLP-IJCNLP*, 2019, pp. 5725–5731.
[69] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4996–5001.
[70] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NIPS*, 2019, pp. 8026–8037.
[71] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*. [Online]. Available: http://arxiv.org/abs/1910.03771
[72] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Proc. COMPLEX Netw.*, 2019, pp. 928–940.
[73] M. A. Di Gangi, G. Lo Bosco, and G. Pilato, "Effectiveness of data-driven induction of semantic spaces and traditional classifiers for sarcasm detection," *Natural Lang. Eng.*, vol. 25, no. 2, pp. 257–285, Mar. 2019.
[74] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS ONE*, vol. 1, Aug. 2020, Art. no. e0237861.



FRANCISCO RODRÍGUEZ-SÁNCHEZ received the B.Sc. and M.Sc. degrees in telecommunications engineering from the Technical University of Cartagena, and the M.Sc. degree in natural language processing (NLP) from UNED, where he is currently pursuing the Ph.D. degree. His research interests include algorithms and NLP techniques to detect hate speech and sexism online.

JORGE CARRILLO-DE-ALBORNOZ received the master's degree in computer science research and the Ph.D. degree in natural language processing. He is currently a Teaching Assistant with UNED. He has participated in several national and EU funded research projects and published over 30 articles in important conferences and journals in the area. His research interests are in discovering and representing sentiments and emotions in text and automatically monitoring social networks for tracking and detecting sensible information about companies and brands.

LAURA PLAZA is currently an Associate Professor with UNED and a Researcher with IR & NLP Group, UNED. She has authored or coauthored in more than 40 international journals and conferences. She has participated in different funded projects and worked in several international companies. Her research includes different fields of natural language processing, including summarization, word sense disambiguation, information retrieval and sentiment analysis, with an emphasis on semantic approaches in the biomedical domain.