Frenda, S.; Ghanem, B.; Montes-Y-Gómez, M.; Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. Journal of Intelligent & Fuzzy Systems. 36(5):4743-4752. https://doi.org/10.3233/JIFS-179023

Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter

Simona Frenda[a,b,*], Bilal Ghanem[b], Manuel Montes-y-Gómez[c] and Paolo Rosso[b]
[a]*Dipartimento di Informatica, Università degli Studi di Torino, Italy*
[b]*PRHLT Center, Universitat Politècnica de València, Spain*
[c]*Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico*

Abstract. Patriarchal behavior, such as other social habits, has been transferred online, appearing as misogynistic and sexist comments, posts or tweets. This online hate speech against women has serious consequences in real life, and recently, various legal cases have arisen against social platforms that scarcely block the spread of hate messages towards individuals. In this difficult context, this paper presents an approach that is able to detect the two sides of patriarchal behavior, misogyny and sexism, analyzing three collections of English tweets, and obtaining promising results.

Keywords: Misogyny detection, sexism detection, linguistic analysis

1. Introduction

Recently, the supervision of user-generated contents has turned into a necessity for Internet companies because of the uncontrolled spread and permanence of hate messages. In this situation, the paradox of Internet arose: while the net could have been a way to guarantee free speech, now social platforms have to protect their users from abusive contents monitoring and censoring them.

The spread of hate ideas online amplifies social misbehavior, supporting and inciting hate crimes in the real world. For instance, the correlation between the increase of misogynistic behavior online and the number of rapes per state in USA is highlighted in [13]. Moreover, a new dangerous social trend is to record with cameras the criminal actions that will be published online, enduring in the web and facilitating the continue extortion of the victims [2].

The consequences of hate speech online are psychological not only for the victims but also for the readers. For instance, in 2014 the workers of the website of the Jezebel companies wrote an open letter to the parent company explaining which are the professional and the emotional costs of daily exposure to violent rape contents in the site comment section [17].

These facts motivate political ventures and stimulate the research attention, especially in the natural language processing (NLP) field. In particular, the shared task about misogyny detection proposed at IberEval[1] is inspirational to this work. Actually, hate against women is a complex topic that involves cultural or traditional customs, that, with difficulty, could be changed. In fact, patriarchal behavior is part of the majority of cultures and, in spite of all the feminist revolutions, the equality of rights is far from being reached.

[1]https://sites.google.com/view/ibereval-2018

Considering this framework, this study aims to analyze the differences and analogies between two aspects of online hate speech against women: misogyny and sexist behavior. For this purpose, the same machine learning approach is used to detect automatically misogynistic and sexist tweets against women in the English language. Misogyny etymologically comes from the Greek word *misoginìa* compound of *miso-* ("to hate") and *-gyne* ("woman") and, as Julie Bindel[2] said, a misogynistic utterance is always sexist:

> i) *Is there a thing in the constitution about women shutting the hell up? COME ON! #shutthehellupwomen*[3].

Indeed, sexism is to discriminate or judge someone (especially women) on the basis of gender, such as in:

> ii) *@jackheathh I'm not sexist but women drivers are bad and when i mean bad I mean BAD*[4].

However, as Kate Manne [20] said, misogyny and sexism work hand-in-hand to uphold patriarchal social relations[5].

Taking into account these definitions of misogyny and sexism, two research questions arose:

RQ1 Could sexist utterances discriminating women be indicative of misogynistic attitude and, then, considered as hate attacks against women?

and, therefore,

RQ2 Is it possible to approach automatically these two social misbehaviors as two sides of the patriarchal mentality?

For this purpose, the proposed system analyzes two corpora containing misogynistic tweets [8, 9] and a corpus containing sexist tweets [25], using the same set of features. In particular, stylistic features and lexicons modeled on topics and semantic information are used. These lexicons have been built to explore the traits of verbal aggression to women.

This study is an initial work that aims at answering the above-mentioned questions, thus exploring the linguistic analogies and differences between sexism and misogyny from a computational perspective.

The main contributions of this work could be summarized as follows:

1. to approach for the first time the automatic detection of misogyny and sexism against women using the same computational approach;
2. to investigate linguistic analogies and differences between sexism and misogyny from a computational point of view;
3. to examine the usefulness of stylistic and lexical features for hate speech online against women.

The rest of the paper is structured as follows. The next section outlines the related work. Section 3 describes the used corpora. Section 4 presents the approach for misogyny and sexism against women detection, reporting the experiments carried out and discussing the obtained results. Finally, Sections 5 and 6 discuss the analysis and draw some conclusions.

2. Related work

The majority of related NLP research is focused especially on abusive language detection. Actually, the necessity to monitor general hate speech encourages list-based techniques that, recognizing profanities, flag the texts as aggressive. However, the authors of [24] demonstrated that the simple use of black-lists records poor performances. This result is due, principally, to misspellings, hate disguised as humour and the contextual nature of profanity. Indeed, the profanities usually are used as simple exclamations.

Therefore, recent researches address efforts to deeply investigate abusive language and the way it is manifested in social media. For instance, in [5] the authors investigated the functional linguistic variations between racist and sexist tweets in the corpus proposed by the authors of [25], discovering that sexist tweets tend to be more interactive and attitudinal than racist ones, addressed principally to persuasion.

On this same corpus, in [14] the authors proposed a deep learning system that assigns each tweet to one of the four categories (racism, sexism, both racism and sexism and non-abusive language). They compared the performance of a Convolutional Neural Network system (with a F1-score of 0.78) with the Logistic Regression classifier based on characters n-grams (with a F1-score of 0.74) used by the authors of [25].

[2] Julie Bindel is a freelance journalist, political activist, and a founder of Justice for Women.

[3] Tweet extracted from the misogynistic corpus of the proposed shared task at IberEval 2018.

[4] Tweet extracted from the sexist corpus proposed by [25].

[5] https://www.vox.com/identities/2017/12/5/16705284/metoo-weinstein-misogyny-trump-sexism

Our work focuses mainly on the sexism category of this corpus (see Section 3).

With respect to sexism against women, the authors of [12, 19] investigated typical aspects of virtual life such as interactivity and anonymity which minimize the authority and the inhibition of the user facilitating sexism attacks. These phenomena are visible especially in the video games context, as demonstrated in the survey proposed by the authors of [11], where social dominance and masculine norms are present and women need to comply with them.

Regarding misogyny, to our knowledge, [1] is the first study to have faced the problem of its automatic identification in Twitter. The authors compared the performance of different supervised approaches using word embeddings, stylistic and syntactic features. The results revealed that the best machine learning approach for misogyny classification is the linear Support Vector Machine (SVM) classifier with an accuracy of 77%.

Machine learning based approaches are the techniques more used in hate speech detection [4, 7, 18, 21]. Actually, they allow researchers to explore closely this issue, exploiting various features such as textual [6] and syntactical aspects [3] or semantic and sentiment information [15, 21, 23].

Although this work is inspired by these previous researches, its main scope is to understand the analogies and differences at computational level of the automatic detection of misogyny and sexism against women online.

3. Datasets

This study is based on three corpora of English tweets: two corpora about misogyny proposed as training sets by the organizers of the automatic misogyny identification (AMI) shared tasks at IberEval 2018[6] [8] and EvalIta 2018[7] [9], and one corpus about sexism collected by the authors of [25].

3.1. Misogynist corpora

Although the organizers of the AMI shared task provided Spanish and English collections at IberEval 2018 and Italian and English corpora at EvalIta 2018, this work is focused only on English language. The idea, in fact, is to compare the results obtained on misogyny corpora with the ones obtained on sexist English corpus.

Table 1
Statistics of the training datasets provided by the AMI organizers at IberEval and EvalIta 2018

Dataset	*Misogynistic*	*Non-Misogynistic*	*Total*
IberEval_2018	1,568	1,683	3,251
EvalIta_2018	1,785	2,215	4,000

Table 2
Statistics of the collection selected from the original SRW

Sexist	*Non-Sexist*	*Total*
2,503	2,503	5,006

In the context of the AMI shared task, the organizers asked participants to detect firstly if the message is misogynistic, and secondly to classify the target (individual or not), and the category of misogyny according to the classes proposed in [22]: stereotype and objectification, dominance, derailing, sexual harassment and threats of violence, and discredit. However, for the purpose of this study, this work focuses only on misogyny detection.

The tweets were collected using keywords and hashtags regarding harassments and attacks on women [1]. Table 1 shows some statistics of AMI's datasets.

3.2. Sexist corpus

The third corpus used in this research is available online (NAACL_SRW_2016_tweets[8]). The authors in [25] provide just the ids of the tweets annotated with "sexist", "racist" and "none" labels. Unfortunately some of the tweets were no longer available[9]. Despite a balanced collection of positive and negative samples is not a well-founded representation of real world, an equal number of sexist and non-sexist tweets is selected such as in the balanced AMI collections. Therefore, all the available sexist tweets are chosen and a correspondent number of non-sexist tweets (labeled "none") is randomly selected from the downloaded set. Hereafter we will refer to this corpus as SRW. Table 2 shows the statistics of the dataset.

3.3. Analysis of corpora

Carrying out the analysis of the considered corpora, similar characteristics appear. For this com-

[6] https://amiibereval2018.wordpress.com/

[7] https://amiEvalIta2018.wordpress.com/

[8] https://github.com/ZeerakW/hatespeech

[9] To download the tweets by ids the Twitter API for Python has been used.

Table 3
Analysis of corpora

	AMI_IberEval	*AMI_EvalIta*	*SRW*
Number of tweets	3,251	4,000	5,006
Vocabulary	9,158	10,532	11,966
Number of tokens	55,431	68,573	79,138
Type-token ratio	16.52%	15.36%	15.12%
Average of words	17.05	17.14	15.81

Table 4
Analysis of positive samples in each corpus

	AMI_IberEval	*AMI_EvalIta*	*SRW*
Number of tweets	1,568	1,785	2,503
Vocabulary	5,155	5,932	7,846
Number of tokens	27,477	31,535	44,803
Type-token ratio	18.76%	18.81%	17.51%
Average of words	17.52	17.67	17.90
Swear words	3,176	3,587	1,261
Feminine pronouns	353	344	251
Masculine pronouns	111	80	66

Table 5
Analysis of negative samples in each corpus

	AMI_IberEval	*AMI_EvalIta*	*SRW*
Number of tweets	1,683	2,215	2,503
Vocabulary	6,228	7,133	6,633
Number of tokens	27,954	37,038	34,335
Type-token ratio	22.28%	19.26%	19.32%
Average of words	16.61	16.72	13.72
Swear words	2,087	2,251	600
Feminine pronouns	158	150	88
Masculine pronouns	157	159	117

parison, the size of corpora, vocabulary, lexical richness and the average of words per tweet are taken into account. In particular, the lexical richness of the collections of tweets is calculated by means of the Type-Token Ratio (TTR) that calculates the variation of the lexicon into each corpus. Table 3 summarizes these aspects of the three corpora.

In order to obtain these values, every symbol and punctuation is cleaned off as well as the urls. Considering the important role played by hashtags and mentions (@user) in the tweet context, they are taken into account as tokens. The showed values reveal that the corpora have a similar percentage of lexical diversity with a soft variation of words per tweets. Despite the different length of the corpora, the similar TTR value suggests that the users could use a similar, and probably informal, lexicon in both contexts. To understand better the analogies between positive (i.e., misogynistic and sexist) and negative (i.e., non-misogynistic and non-sexist) samples in all corpora, a linguistic analysis was carried out. Tables 4 and 5 show the obtained values.

In particular, for this analysis, an available online lexicon of English swear words[10] has been used. Moreover, hashtags and mentions are taken into account considering the fact that they could contain also offensive words.

As Tables 4 and 5 show, despite the different number of tweets in the corpora, the TTR and the number of words per positive samples is very similar, differently from values of negative samples.

Focusing on the differences between the values obtained for positive and negative samples, we can see that a richer lexicon is used in non-misogynistic and non-sexist tweets. This factor could be due to the fact that misogynistic and sexist texts mainly contain a substantial number of insults and profanities. Actually, the positive samples aim to offend and hurt the target.

Moreover, a very simple investigation about the presence of masculine pronouns ("he", "his") and feminine pronouns ("she", "her") was carried out. As Tables 4 and 5 show, the positive samples are principally focused on women respect to negative ones. As well as, the obtained values indicate that also sexist tweets talk more about women than men.

Finally, the average of words per tweets reveals that misogynistic and sexist messages are longer than non-misogynistic and non-sexist ones. Indeed, the user tends to justify the negativity of his opinion or underline that his statement is not misogynistic or sexist, such as:

> iii) "*Because femininity is so horrible! @JonnyG313 I'm not sexist but if a dude cries because of a girl in a wedding dress then he has a vagina*"[11].

These analysis reveals some analogies between the tweets annotated with the labels "misogynistic" and "sexist". This similarity is inferred from some examples extracted from the corpora and reported in Table 6. In particular, they target women with the purpose to discredit them and underline male superiority.

4. Proposed approach

The previous analysis and the examples reported in Table 6 indicate the fact that sexist and misogynistic corpora are similar. Moreover, the fact that sexist tweets are more focused on women than men suggests

[10] https://www.cs.cmu.edu/ biglou/resources/

[11] Tweet extracted from SRW corpus.

Table 6
Misogynistic and Sexist examples against women

Corpora	*Tweet*
AMI_IberEval	*What do you call a women that has a brain? Pregnant with a baby boy.* *What's worse than a girl who gives rough handjobs? A feminist.*
AMI_EvalIta	*@Corter_back no I said hope. I hope you women learn your place! #SitDownInTheKitchen* *Who makes the sandwiches at a feminist rally?*
SRW	*RT @boggsy9 Dont ever let women drive, they'll break your arm! #notsexist* *Are you even a real person? @awesomeadanxd I'm not sexist. But Men are superior to women.*

the idea that sexist messages in the most of cases discriminate women. Despite the corpora that we used are representative of a little part of real big data on the web, they help to understand the aspects of hate speech against women.

The problem of the identification of sexist (presumably against women) and misogynistic tweets is addressed as a classification task. On the basis of the performance obtained in the previous work [1], we also employed SVM to detect misogyny in English corpus in our experiments.

The employed classification approach is based on modeled lexicons about specific aspects of online hate speech against women and on stylistic features captured by means of n-grams of words and characters. In particular, we used SVM with the radial basis function kernel (RBF), the parameters of C equal 5 and *gamma* equal 0.1.

To perform the experiments, the tweets are preprocessed by deleting all symbols, emoticons and urls. In order to perform a correct match between the dictionary of the corpora and each lexicon, the lemmatizer provided in the Natural Language Toolkit (NLTK[12]) is applied.

Each tweet is represented as a vector composed of: the weights of n-grams of characters and words calculated with the Term Frequency-Inverse Document Frequency (TF-IDF) measure; and the weights of lexical features calculated with the Information Gain. In order to take into account also the words that are relevant for the classification but they are not in the lexicons, we added in the vector their weights calculated by means of Information Gain.

Finally, the evaluation is performed using 10-fold cross validation to inspect deeply the performance of the used approach.

The following subsections describe the employed features, the experiments carried out and the obtained results.

[12]https://www.nltk.org/

4.1. Linguistic features

As said in the previous section, the employed approach is based on lexical and stylistic features. About lexical features, some lexicons are manually modeled in order to capture specific aspects of aggressive messages against women. In particular, they principally concern femininity, vulgarity, sexuality and human body.

For this purpose, the relevant words of misogynistic and sexist tweets are extracted calculating their weights with the Information Gain measure. As well as, typical linguistic elements of digital writing (such as slangs, abbreviations and hashtags) are taken into account. Along with them, the stylistic aspects are captured by means of n-grams of characters and words. Specifically, the experiments reveal optimal performances with characters from 1 to 7 grams and unigrams, bigrams and trigrams of words. Below, a brief description of these features.

Vulgarity. This lexicon contains vulgarities and offensive adjectives that aim at offending and humiliating, such as: "*whore*", "*slut*", "*slave*" and "*bitch*". In particular the offensive list of words provided by Luis von Ahn's research group[13] is employed in this work and extended manually with 170 words especially focused on attacks to women. In the previous studies [5, 16], the authors underline that some terms are used in many different contexts and often without offensive purposes. In this work, the selection of the words is based on the relevance of the word calculated by means of Information Gain. Considering these weights of the words, the terms such as "*fuck*" are not taken into account. Indeed, these words could be used also as simple exclamation.

Femininity. To match the woman as target of offenses, this lexicon collects about 90 terms rela-

[13]The research group is from Carnegie Mellon's School of Computer Science and provides linguistic resources online.

tive to woman (as "*lady*", "*girl*" or "*pregnant*"). It contains also words that have negative connotation (like "*barby*") and the pronouns such as "*her*", "*she*", "*herself*".

Sexuality. The 290 words collected in this lexicon concern sexual context, perversion and prostitution. Actually, one of the most frequent subjects in hate speech against women is the sexual dominance of man. Some examples from this lexicon are: "*virginity*", "*blowjob*", "*cum*" or "*fingering*".

Human body. In strong connection with sexual lexicon, a set of terms about feminine body has been created. It contains 50 words such as "*pussy*", "*boobs*", "*butt*" and "*legs*".

Hashtag. This list contains about 40 hashtags referring to stereotypes, inferiority of women and sexual harassment, such as #*bitchesaredogs*, #*fuckbitches*, #*keepwomendown*, #*womenaredemons* and #*YesAllWomenBelongInTheKitchen*.

Abbreviations. In particular 50 vulgar typical abbreviations typical of social media languages are collected in this list. Some of the most frequents are *idgaf* ("I don't give a fuck"), *wtf* ("What the Fuck!"), *smh* ("So much hate") and *fkd* ("Fucked").

Character n-grams. The stylistic aspects of the texts are captured by the characters n-grams with a range from 1 to 7 weighted by means of TF-IDF. These features help the system to catch similar patterns in spite of the orthographic errors typical of spontaneous writing in tweets.

Table 7
Some n-grams of characters from the three corpora

AMI_IberEval	*AMI_EvalIta*	*SRW*
' the '	' thi'	' thei'
'#male'	'a bi'	'thr'
'#ye'	'a bit'	' thes'
'our a'	'out o'	'swe'
'#yes'	'to r'	'n an'
'tim '	'a c'	's i '
' ther'	' thou'	' #m'
' a b'	'end '	'hel'
'hea'	' a bi'	' #mk'
'eon '	'hen i'	' #mkr'

Examining the characters n-grams extracted from the datasets, the tweets present similar constructions typical of the informal speech ("*gonn*", "*I m*" , "*yo*"). In addition, words such as "*hoe*" and "*bitc*" are located. Table 8 reports some characters n-grams with the highest Information Gain values extracted from the datasets.

Bag and sequences of words. As said above the unigrams, bigrams and trigrams perform well. Specifically, bigrams and trigrams help the system to recognize specific syntactical and lexical patterns of hate speech against women which are difficult to be captured taking into account only the lists of words. Indeed, as well as character n-grams, also word n-grams preserve the syntactical order of the grams differently from the an approach based exclusively on lists of words. Some examples of significant bigrams and trigrams are reported in Table 7. As Table 7 shows, the subject of the dataset and the women as target emerge from a simple analysis of the most frequent bigrams and trigrams.

The next section describes the experiments carried out and the obtained results.

Table 8
Some examples from the most frequent n-grams of words

Bigrams			**Trigrams**		
AMI_IberEval	*AMI_EvalIta*	*SRW*	*AMI_IberEval*	*AMI_EvalIta*	*SRW*
('a', 'bitch')	('a', 'woman')	('but', 'women')	('Fuck', 'off', 'you')	('Shut', 'fuck', 'up')	('I', 'am', 'sexist')
('a', 'girl')	('a', 'hoe')	('sexist', 'but')	('What', 'the', 'difference')	('WomenSuck', 'don', 't')	('I', 'm', 'sexist')
('a', 'whore')	('a', 'bitch')	('girls', 'are')	('a', 'ass', 'bitch')	('a', 'stupid', 'bitch')	('I', 'not', 'sexist')
('bitch', 'I')	('women', 'are')	('women', 'are')	('a', 'good', 'girl')	('a', 'whore', 'you')	('Not', 'sexist', 'but')
('your', 'ass')	('she', 's')	('but', 'girls')	('a', 'hoe', 'I')	('don', 't', 'get')	('but', 'I', 'hate')
('a', 'woman')	('re', 'a')	('women', 'should')	('a', 'hoe', 'bitch')	('don', 't', 'have')	('but', 'women', 'are')
('a', 'hoe')	('a', 'whore')	('girls', 'should')	('a', 'stupid', 'bitch')	('don', 't', 'know')	('call', 'me', 'sexist')
('stupid', 'bitch')	('stupid', 'bitch')	('no', 'sexist')	('a', 'whore', 'you')	('fuck', 'up', 'you')	('sexist', 'I', 'hate')
('she', 's')	('a', 'girl')	('but', 'female')	('bitch', 'suck', 'my')	('is', 'a', 'bitch')	('sexist', 'I', 'just')
('bitch', 'you')	('Women', 'are')	('when', 'women')	('on', 'my', 'dick')	('is', 'a', 'cunt')	('sexist', 'but', 'female')
('my', 'dick')	('I', 'hate')	('women', 'can')	('re', 'a', 'bitch')	('is', 'a', 'whore')	('sexist', 'but', 'girls')
('ass', 'bitch')	('to', 'rape')	('promo', 'girls')	('re', 'a', 'whore')	('women', 'are', 'stupid')	('sexist', 'but', 'hate')
('you', 'stupid')	('Fuck', 'you')	('all', 'female')	('the', 'difference', 'between')	('you', 'stupid', 'bitch')	('sexist', 'but', 'women')

Table 9
Results based on accuracy obtained for each corpus

	AMI_IberEval	*AMI_ EvalIta*	*SRW*
baseline	*0.7507*	*0.7834*	*0.8836*
Bag and sequences of words	0.7453	0.7860	**0.8932**
Characters n-grams	**0.7544**	**0.7877**	0.8711
Lexicons	0.6994	0.7347	0.7347
All	**0.7605**	**0.7947**	**0.8937**

4.2. Experiments and evaluation

For the comparison of the experiments carried out with the SVM classifier, a simple baseline is obtained using bag of words weighted by means of TF-IDF. Considering the balanced nature of the datasets, the measure of evaluation employed is the accuracy and all the experiments are carried out in the 10-fold cross scenario. The obtained results for each dataset are showed in Table 9. In particular, the classification is carried out firstly taking into consideration the individual features, and secondly considering the combination of the considered features.

As Table 9 shows, stylistic features in general perform well in spite of the limited increase respect to the baseline. Indeed, the combination of all the features seems to work better, reaching higher results than baseline. Moreover, despite the lexicons do not overcome the baseline, they seem to perform similarly in all the corpora. However, in Table 9 a main difference is evident: in the misogynistic corpora the character n-grams appear the best performing feature, while in the sexist corpus the best one is the combination of bag and sequences of words. Looking at Table 8 and Table 7, it is possible to hypothesize that the character n-grams are important in a context where the vocabulary is composed of a great amount of insults. While in the sexist corpus where insults are less frequent, the co-occurrence of words seem to be important patterns for the classification.

In order to understand better how our classifier works, Table 4.2 reports some examples correctly predicted using the combination of all the features.

4.3. Analysis of lexicons

Considering the results of the previous section and mainly the similar performance obtained in all the corpora by the lexical features, a linguistic analysis is carried out by means of Information Gain. A resume of the value of lexicons for each corpus is reported in Fig. 1.

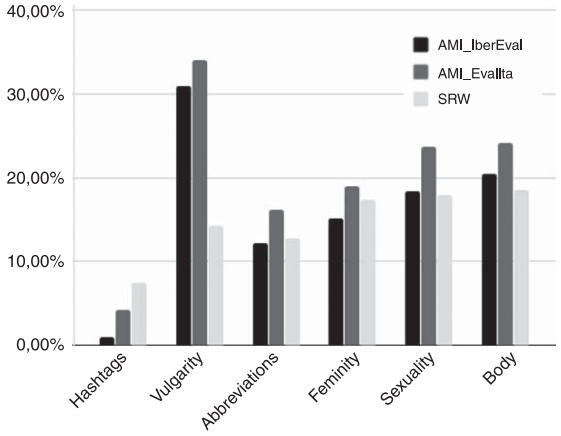

Fig. 1. Analysis of the lexicons by means of Information Gain.

As the figure shows the differences between the corpora are minor, indicating that the employed lexical features are important for all the corpora. In particular, Sexuality and Human Body lexicons play an important role in all the corpora, because they are the main way to discriminate women from men. Also words relative to Femininity have similar relevant distribution in all corpora, as well as the use of abbreviation. Specifically, Table 11 reports some examples concerning these main analogies.

Despite these similarities, a strong difference is the number of swear words (evident also in Table 4 and Table 5). Indeed, Fig. 1 shows that the list of profanities and offensiveness has a higher impact especially on the misogynistic corpora, suggesting that the tweets of these corpora aim to offend more than the sexist tweets. However, these results do not hide the fact that sexist tweets are usually addressed to women and that sexist discrimination, as the various examples show, could contribute to offend women.

4.4. Error analysis

To understand what are the causes that could hinder the correct classification of misogynistic and sexist tweets against women, an error analysis is carried out. Despite the natural differences such as the schema of annotation or the keywords used to collect the tweets, the misclassified instances are similar. For instance, the lack of knowledge of the world is an evident obstacle for the system. One of these cases is in the sexist corpus. In fact, some tweets, such as:

Table 10
Correctly predicted tweets with a combination of features: n-grams (underlined) and words from lexicons (in italics)

Corpora	*Tweet*
AMI_IberEval	@EbersohlMyra @Lolomonet @mandee_boo @_morganmariaa *Bitch* you look like a dog Control, I own your ass now *bitch* This *stupid* Bitch Lied. https://t.co/KOeOdaibX7
AMI_EvalIta	Too many *women* confuse their hearts with their *vaginas* The more *WOMEN* in the workplace means *LESS WOMEN* IN THE KITCHEN. Like and retweet if you see this as a big problem #ilovefood #homecookedm @JWMofficial: You're wearing yoga pants to show off your ass. don't give me this *bull* shit 'they're comfy!' we all know the truth *ladies #ASS #ASS #ASS*
SRW	@mannythemenace Call me sexist, but I hate female receptionists. They always give me attitude like I'm the reason their *bf* broke up with 'em RT @RobDurbinn I'm not *sexist* but *women* rappers are a bigger joke than the WNBA @terrencewoods I'm not *sexist* but *girls* our age that drive is super scary

Table 11
Misogynistic and sexist tweets containing words from Sexuality (underlined) and Femininity (in italics) lexicons

Corpora	*Tweet*
AMI_IberEval	Me trying to flirt- You have really nice eyebrows... I'd like to cum on them to see if they wash off RT @KGJump12: Sometimes I want a *girlfriend*, but then I quickly remember how i hate *women*
AMI_EvalIta	Love a bitch for what ? all she good for is sucking dick! They've made it almost impossible for Men to be dominant, in the *matriarchal* western society. fuckfeminism
SRW	RT @of_The_Guild It really pisses me off when anime girls don't have big boobs #NotSexist - A Misogynist @parody_guy A woman wants *her* man to treat *her* like a *princess* to the world and fuck *her* like a whore. - Someone

iv) *"@SydneyEditor1: Not enough Lemon in their lemon tart #MKR woops". Plenty of tart though*

talks about the Australian competitive cooking game show: My Kitchen Rules (MKR). Actually, these comments refer to participants or some events in the show that are difficult to understand without the context. The authors in [25] specified that they used the hashtags relative to MKR to collect the data, because often the tweets containing #mkr prompt sexist issues directed at the female participants.

Among the wrong predictions of the sexist corpus, few tweets like the one below have not been predicted as sexist because the target of these messages is not the woman but the man:

v) *"@philippenis @DannyVelasco I made up nothing. A stoner said that and you are too ignorant of world events to know he's wrong. #sorryitsaboy".*

This kind of error was foreseeable considering the nature of the corpus. However, for the aim of this work, this "error" suggests that the system works well.

Another problem arose from the lack of context concerning the presence of urls in the tweets. Indeed, some of them refer to external information conveyed by the links:

vi) *@amberhasalamb Can you explain why this is wrong? http://t.co/pTkwk45P9P.*

In particular, this URL recalls another tweet which deals with the common idea that when women are angry it is supposedly due to the lack of sex in their lives. For this reason, conveyed by the information in the link, the tweet could be annotated as sexist. However, for the purpose of this research, the proposed approach does not take into consideration the url content.

Finally, there are some tweets that describe a sexist or misogynistic situation, such as the following tweet:

vii) *@AJEnglish It is Muslim Jihad culture to rape yagidi in Iraq,Christian & Hindu in Pak.purchase poor Muslim girl for sex slave.*

This kind of texts are not classified as aggressive against women because they are not addressed specifically to women, but they describe an event or situation. Moreover, this specific tweet shows compassion towards and not attack against women, thus the system classifies well for our proposal.

5. Discussion

Generally, the collections of data extracted from the web are a little representation of all the texts that are published daily on the web by millions of people, and thus a little representation of their opinions. Focusing on the corpora here analyzed, it is possible to notice that the texts that are annotated as sexist and that theoretically discriminate male as well as women, really they tend to offend mainly the women. This observation is confirmed not only by corpora analysis but also by the analysis of n-grams of characters and words extracted from the texts. Moreover, the fact that the implemented system oriented to identify the aggressive texts against women works well in the analyzed sexist corpus as well as in the misogynistic corpora reveals that these kind of collections contain similar texts. Therefore, answering the first research question, it seems that the discrimination of women could be a signal of hate attitude against women. Could the sexist "common way of thinking" be considered innocent? The computational experiments presented in this work seem to affirm that in general sexist opinions hide a sentiment of hate and, in this particular case, a misogynistic attitude. Despite sexist humour is commonly considered guiltless, various studies affirm the contrary. For instance, [10] underlines that sexist jokes are experienced in the same way as misogynistic assertion. Furthermore, sexist jokes could contribute to make sexism or misogyny like a norm although also this kind of jokes hurt the target.

6. Conclusions and further work

This paper presents a novel study on online hate speech against women focusing on the differences and analogies between misogyny and sexism. In particular, this study focuses on Twitter data analysis.

On the basis of the analysis that we did, these two phenomena reported similar characteristics. Moreover, it could be interesting to perform new experiments also to identify discrimination against not only women but also men in the same sexist corpus in order to notice if the same sentiment of hate emerges for other targets.

However, the error analysis does not hide the limitations of a lexicon-based approach. Therefore, as future work, we aim at exploring the semantic dimension by means of deep learning techniques.

Moreover, we aim to see if some findings will be also present in other corpora and languages, for instance in the English and Spanich corpora of the hatEval multilingual shared tasks at SemEval 2019[14] (whose targets of hate speech will be women and also immigrants).

Considering that the humour in some case could be used also as a mask to disguise the negative opinion as well as sexist, racist or misogynistic assertions, a future study will be concentrated on the role of the humour in hate speech automatic detection.

Starting from two principal research questions, this work explored the analogies and differences between two social misbehaviors principally targeting women with the aim to offend, discriminate and hurt them. From a computational perspective this study aimed at exploiting these analogies to identify hate speech against women. Promising results seem to confirm the initial hypothesis, which sees sexist and misogynistic attitudes as expressions of the patriarchal mentality.

Acknowledgments

The work of Simona Frenda and Paolo Rosso was partially funded by the Spanish MINECO under the research project SomEMBED (TIN2015-71147-C2-1-P). We also thank the support of CONACYT-Mexico (project FC-2410).

References

[1] M. Anzovino, E. Fersini and P. Rosso, Automatic Identification and Classification of Misogynistic Language on Twitter, *Proc 23rd International Conference on Applications of Natural Language to Information Systems, NLDB-2018*, Springer-Verlag, LNCS 10859, 2018, pp. 57–64.

[2] C. Buni and S. Chemaly, The Unsafety Net: How Social Media Turned Against Women, The Atlantic, https://www.theatlantic.com/technology/archive/2014/10/theunsafety-net-how-social-media-turned-against-women/381261/, 09/10-2014.

[3] P. Burnap and M.L. Williams, Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making, *Internet, Policy and Politics*, Oxford, UK, 2014.

[4] P. Burnap, O.F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan and L. Sloan, Detecting tension in online communities with computational Twitter analysis, *Technological Forecasting and Social Change* **95** (2015), 96–108.

[14] http://alt.qcri.org/semeval2019/index.php?id=tasks

[5] I. Clarke and J. Grieve, Dimensions of abusive language on twitter, *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 1–10.

[6] Y. Chen, Y. Zhou, S. Zhu and H. Xu, Detecting offensive language in social media to protect adolescent online safety, *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, Amsterdam, Netherlands, IEEE, 2012, pp. 71–80.

[7] H.J. Escalante, E. Villatoro-Tello, S.E. Garza Villareal, A.P. López-Monroy, M. Montes-y-Gómez and L. Villaseñor-Pineda, Early detection of deception and aggressiveness using profilebased representations, *Expert Systems with Applications* **89** (2017), 99–111.

[8] E. Fersini, M. Anzovino and P. Rosso, Overview of the Task on Automatic Misogyny Identification at IBEREVAL, *CEUR Workshop Proceedings 2150*, Seville, Spain, 2018.

[9] E. Fersini, D. Nozza and P. Rosso, Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI), *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy, 2018.

[10] T.E. Ford and C.F. Boxer, Sexist humor in the workplace: A case of subtle harassment, *Insidious Workplace Behavior*, Routledge, 2011, pp. 203–234.

[11] J. Fox and W.Y. Tang, Sexism in online video games: The role of conformity to masculine norms and social dominance orientation, *Computers in Human Behavior* **33** (2014), 314–320.

[12] J. Fox, C. Cruz and J.Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, *Computers in Human Behavior* **52** (2015), 436–442.

[13] 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, *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*, 2014.

[14] B. Gambäck and U.K. Sikdar, Using convolutional neural networks to classify hate-speech, *Proceedings of the First Workshop on Abusive Language Online* 2017.

[15] N.D. Gitari, Z. Zuping, H. Damien and J. Long, A lexiconbased approach for hate speech detection, *International Journal of Multimedia and Ubiquitous Engineering* **10**(4) (2015), 215–230.

[16] S. Hewitt, T. Tiropanis and C. Bokhove, The problem of identifying misogynist language on Twitter (and other online social spaces), *Proceedings of the 8th ACM Conference on Web Science*, 2016, pp. 333–335.

[17] Jezebel, What Gawker Media Is Doing About Our Rape Gif Problem, Jezebel Website, https://jezebel.com/what-gawkermedia-is-doing-about-our-rape-gif-problem-1620742504,11/08/2014, 2014.

[18] R. Justo, T. Corcoran, S.M. Lukin, M. Walker and M.I. Torres, Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web, *Knowledge-Based Systems*, 2014.

[19] N. Lapidot-Lefler and A. Barak, Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition, *Computers in Human Behavior* **28**(2) (2012), 434–443.

[20] K. Manne, *Down Girl: The Logic of Misogyny,* Oxford Scholarship Online, 2017.

[21] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, Abusive language detection in online user content, *Proceedings of the 25th International Conference on World Wide Web*, Geneva, Switzerland, 2016, pp. 145–153.

[22] B. Poland, *Haters: Harassment, Abuse, and Violence Online*, University of Nebraska Press, Lincoln, 2016.

[23] N.S. Samghabadi, S. Maharjan, A. Sprague, R. Diaz-Sprague and T. Solorio, Detecting nastiness in social media, *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, Canada, 2017, pp. 63–72. Association for Computational Linguistics.

[24] S. Sood, J. Antin and E. Churchill, Profanity use in online communities, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, 2012, pp. 1481–1490.

[25] Z. Waseem and D. Hovy, Hateful symbols or hateful people, *Predictive Features for Hate Speech Detection on Twitter. HLT-NAACL*, 2016.