He said "who's gonna take care of your children when you are at ACL?": Reported Sexist Acts are Not Sexist

Patricia Chiril[1], Véronique Moriceau[1], Farah Benamara[1]
Alda Mari[2], Gloria Origgi[2], Marlène Coulomb-Gully[3]
(1) IRIT, Université de Toulouse, Université Toulouse III - UPS, France
{firstname.lastname}@irit.fr
(2) Institut Jean Nicod, CNRS, ENS, EHESS, Paris, France
{firstname.lastname}@ens.fr
(3) LERASS, Université de Toulouse, UT2J, France
marlene.coulomb@univ-tlse2.fr

Abstract

In a context of offensive content mediation on social media now regulated by European laws, it is important not only to be able to automatically detect sexist content but also to identify if a message with a sexist content is really sexist or is a story of sexism experienced by a woman. We propose: (1) a new characterization of sexist content inspired by speech acts theory and discourse analysis studies, (2) the first French dataset annotated for sexism detection, and (3) a set of deep learning experiments trained on top of a combination of several tweet's vectorial representations (word embeddings, linguistic features, and various generalization strategies). Our results are encouraging and constitute a first step towards offensive content moderation.

1 Introduction

Sexism is prejudice or discrimination based on a person's gender. It is based on the belief that one sex or gender is superior to another. It can take several forms from sexist remarks, gestures, behaviours, practices, insults to rape or murder. Sexist hate speech is a message of inferiority usually directed against women at least in part because they are women, some authors refer to it as: "words that wound" (Matsuda et al., 1993; Waldron, 2012; Delgado et al., 2015). As defined by the Council of Europe, "The aim of sexist hate speech is to humiliate or objectify women, to undervalue their skills and opinions, to destroy their reputation, to make them feel vulnerable and fearful, and to control and punish them for not following a certain behaviour"[1]. Its psychological, emotional and/or physical impacts can be severe. In several countries, sexist behaviours are now prohibited. See for example the French law of 27 January 2017 related to equality and citizenship, where penalties due to discrimination are doubled (sexism is now considered as an aggravating factor), law that extends to the internet and social media.

Although overall misogyny and sexism share the common purpose of maintaining or restoring a patriarchal social order, Manne (2017) illustrates the contrast between the two ideologies. A sexist ideology (which often "consists of assumptions, beliefs, theories, stereotypes and broader cultural narratives that represent men and women") will tend to discriminate between men and women and has the role of justifying these norms via an ideology that involves believing in men's superiority in highly prestigious domains (i.e., represents the "justificatory" branch of a patriarchal order). A misogynistic ideology does not necessarily rely on people's beliefs, values, and theories, and can be seen as a mechanism that has the role of upholding the social norms of patriarchies (i.e., represents the "law enforcement" branch of a patriarchal order) by differentiating between good women and bad women and punishing those who take (or attempt to take) a man's place in society. Considering these definitions, misogyny is a type of sexism. In this paper, as we target French sexist messages detection, we consider sexism in its common French usage, i.e. discrimination or hate speech against women.

Social media and web platforms have offered a large space to sexist hate speech (in France, 10% of sexist abuses come from social media (Bousquet et al., 2019)) but also allow to share stories of sexism experienced by women (see "The Everyday Sexism Project"[2] available in many languages, "Paye ta shnek"[3] in French, or hashtags such as #metoo or #balancetonporc). In this context, it is important to automatically detect sexist messages

[1] https://rm.coe.int/1680651592

[2] https://everydaysexism.com/

[3] https://payetashnek.tumblr.com/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4055–4066

on social platforms and possibly to prevent the wide-spreading of gender stereotypes, especially towards young people, which is a first step towards offensive content moderation (see the recommendations of the European commission (COM, 2017). However, we believe that it is important not only to be able to automatically detect messages with a sexist content but also to distinguish between real sexist messages that are addressed to a woman or describing a woman or women in general (e.g., *The goalkeeper has no merit in stopping this pregnant woman shooting*), and messages which relate sexism experiences (e.g., *He said "who's gonna take care of your children when you are at ACL?"*). Indeed, whereas messages could be reported and moderated in the first case as recommended by European laws, messages relating sexism experiences should not be moderated.

As far as we are aware, the distinction between reports/denunciations of sexism experience and real sexist messages has not been addressed. Previous work considers sexism either as a type of hate speech, along with racism, homophobia, or hate speech against immigrants (Waseem and Hovy, 2016; Golbeck et al., 2017; Davidson et al., 2017; Basile et al., 2019; Schrading et al., 2015) or study it as such. In this latter case, detection is casted as a binary classification problem (sexist vs. non-sexist) or a multi-label classification by identifying the type of sexist behaviours (Jha and Mamidi, 2017; Sharifirad et al., 2018; Fersini et al., 2018b; Karlekar and Bansal, 2018; Parikh et al., 2019). English is dominant, although Italian and Spanish have already been studied (see the IberEval 2018 (Fersini et al., 2018b), EvalIta 2018 (Fersini et al., 2018a) and HateEval 2019 (Basile et al., 2019) shared tasks).

This paper proposes the first approach to detect different types of reports/denunciations of sexism experiences in French tweets, based on their impact on the target. Our contributions are:
(1) A novel characterization of sexist content-force relation inspired by speech acts theory (Austin, 1962) **and discourse studies in gender** (Lazar, 2007; Mills, 2008). We distinguish different types of sexist content depending on the impact on the addressee (called 'perlocutionary force'): sexist hate speech *directly addressed* to a target, sexist *descriptive assertions* not addressed to the target, or *reported assertions* that relate a story of sexism experienced by a woman. This is presented in Section 3. Our guiding hypothesis is that indirect acts establish a distancing effect with the reported content and are thus less committal on behalf of the addressee (Giannakidou and Mari, 2021). Our take on the issue is language-driven: reported speech is indirect, and it does not discursively involve a call on the addressee to endorse the content of the act.
(2) The first French dataset of about $12,000$ **tweets annotated for sexism detection** according to this new characterization[4]. Data and manual annotation are described in Section 4.
(3) A set of experiments to detect sexist content in three configurations: binary classification (sexist content vs. non-sexist), three classes (reporting content vs. non-reporting vs. non-sexist), and a cascade classifier (first sexist content and then reporting). We rely on deep learning architectures trained on top of a combination of several tweet's vectorial representations: word embeddings built from different sources, linguistic features, and various generalization strategies to account for sexist stereotypes and the way sexist contents are linguistically expressed (see Section 5). Our results, presented in Section 6, are encouraging and constitute a first step towards automatic sexist content moderation.

2 Related Work

Gender in discourse analysis. Discourse analysis studies have shown that sexism may be expressed at different linguistic granularity levels going from lexical to discursive (Cameron, 1992): e.g., women are often designated through their relationship with men or motherhood (e.g., *A man killed in shooting* vs. *Mother of 2 killed in crash*) or by physical characteristics (e.g., *The journalist who presents the news* vs. *The blonde who presents the news*). Sexism can also be hostile (e.g., *The world would be a better place without women*) or benevolent where messages are subjectively positive, and sexism is expressed in the form of a compliment (e.g., *Many women have a quality of purity that few men have*) (Glick and Fiske, 1996). In communication studies, the analysis of political discourse (Bonnafous, 2003; Coulomb-Gully, 2012), sexist abuse or media discourse (Dai and Xu, 2014; Biscarrat et al., 2016) show that political women presentations are stereotyped: use of physical or clothing character-

[4] https://github.com/patriChiril/An-Annotated-Corpus-for-Sexism-Detection-in-French-Tweets

istics, reference to private life, etc. From a sociological perspective, studies focus on social media contents (tweets) or SMS in order to analyze public opinion on gender-based violence (Purohit et al., 2016) or violence and sexist behaviours (Barak, 2005; Megarry, 2014).

Gender bias in word embeddings. Bolukbasi et al. (2016) have shown that word embeddings trained on news articles exhibit female/male gender stereotypes. Several algorithms have then been proposed to attenuate this bias (Dev and Phillips, 2019) or to make embeddings gender-neutral (Zhao et al., 2018), although Gonen and Goldberg (2019) consider that bias removal techniques are insufficient. Debiased embeddings were used by Park et al. (2018) observing a decrease in sexism detection performance compared to the non-debiased model. To overcome this limitation, Badjatiya et al. (2019) propose neural methods for stereotypical bias removal for hate speech detection (i.e., hateful vs. non-hateful). They first identify a set of bias sensitive words, then mitigate their impact by replacing them with their POS, NER tags, K-nearest neighbours and hypernyms obtained via WordNet.

Automatic sexism detection. To our knowledge, the automatic detection of sexist messages currently deals only with English, Italian and Spanish. For example in the *Automatic Misogyny Identification* (AMI) shared task at IberEval and EvalIta 2018, the tasks consisted in detecting sexist tweets and then identifying the type of sexist behaviour according to a taxonomy defined by (Anzovino et al., 2018): discredit, stereotype, objectification, sexual harassment, threat of violence, dominance and derailing. Most participants used SVM models and ensemble of classifiers for both tasks with features such as n-grams and opinions (Fersini et al., 2018b). These datasets have also been used in the *Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter* shared task at SemEval 2019. Best results were obtained with an SVM model using sentence embeddings as features (Indurthi et al., 2019).

There are also a few notable neural network techniques. Jha and Mamidi (2017) employ an LSTM model to classify messages as: benevolent, hostile and non-sexist. Zhang and Luo (2018) implement two deep neural network models (CNN + Gated Recurrent Unit layer and CNN + modified CNN layers for feature extraction) in order to classify social media texts as racist, sexist, or non-hateful. Karlekar and Bansal (2018) use a single-label CNN-LSTM model with character-level embeddings to classify three forms of sexual harassment: commenting, ogling/staring, and touching/groping. Sharifirad et al. (2018) focus on diverse forms of sexist harassment (indirect, information threat, sexual, physical) using LSTM and CNN on augmented dataset obtained via ConceptNet *is-a* relationships and Wikidata. Finally, (Parikh et al., 2019) consider messages of sexism experienced by women in the "Everyday Sexism Project" web site and classify them according to 23 non mutually exclusive categories using LSTM, CNN, CNN-LSTM and BERT models trained on top of several distributional representations (character, subwords, words and sentence) along with additional linguistic features.

In this paper, we propose different deep learning architectures to detect reporting of sexist acts and, more importantly, distinguishing them from real sexist messages. We explore BERT contextualized word embeddings trained from several sources (tweets, Wikipedia) complemented with both linguistic features and generalization strategies. These strategies are designed to force the classifier to learn from generalized concepts rather than words, which may be rare in the corpus. We, therefore, adopt several replacement combinations based on a taxonomy of stereotyped gendered words coupled with additional sexist vocabularies extending Badjatiya et al. (2017) approach designed for hate speech detection to sexism content detection.

3 Characterizing Sexist Content

Propositional content can be introduced in discourse by acts of varying forces (Austin, 1962): it can be asserted (e.g., *Paul is cleaning up his room*), questioned (e.g., *Is Paul cleaning up his room*?), or asked to be performed as with imperatives (e.g., *Paul, clean up your room*!). In philosophy of language, on the one hand, and feminist philosophy on the other, speech acts have already been advocated in a variety of manners. Most accounts however either focus on the type of act (assault-like, propaganda, authoritative, etc.) that derogatory language performs (Langton, 2012; Bianchi, 2014) or concentrate on the analytical level at which the derogatory content is interpreted, whether it provides meaning at the level of the presupposition (or more largely non at-issue content (Potts, 2005)) or of the assertion (Cepollaro, 2015).

We have chosen to distinguish cases where the

addressee is directly addressed from those in which she is not, as done in hate speech analysis. For example, Waseem et al. (2017) and ElSherief et al. (2018) consider that directed hate speech is explicitly directed at a person while generalized hate speech targets a group. For (Ousidhoum et al., 2019), a hateful tweet is direct when the target is explicitly named, or indirect when "less easily discernible". Unlike these approaches and the definitions of target used in (Basile et al., 2019; Fersini et al., 2018a), we do not consider the number of targets of a sexist message (it can indifferently be a woman, a group of women or all women) but rather distinguish the target from the addressee. Our use of the notions of directness and indirectness are also transverse to the ones used in (Lazar, 2007; Chew and Kelley-Chew, 2007) or (Mills, 2008), who resort to the label indirectness for subtle forms of sexism that perpetuate gender stereotypes through humor, presuppositions, metaphors, etc.

We newly consider three different stages in the scale of 'directedness' of an assertion: assertions directed to the addressee, descriptive assertions not directed to a particular addressee and reported assertions. All these three types of acts can contain subtle and non-subtle sexist content. The main goal of our classification is thus to focus on the impact of the content by resorting to the force of the act and not only to its content.

Sexist content in **directed assertions** is explicitly addressed at a target, but contrary to other approaches cited above, the target can be a woman, a group of women or all women. Across the different classifications of speech acts (Portner, 2018), 'direct' speech acts such as imperatives are addressee-oriented and they require that the addressee performs an action (responding (with questions) or acting (with imperatives)). Indirect speech acts are not addressee-oriented. Assertions themselves can be direct or indirect. They are direct when they are in the second person ('you'), as shown in (1) and (2) (linguistic clues are underlined)[5]. They require that the addressee be committed to the truthfulness of their content. Since a direct sexist assertion is a type of speech act that immediately involves the addressee and triggers a request of commitment, direct assertions of sexism have been ranked as the most prominent expressions of sexism with a greater impact on the victim. Most prominently, with assertions, directedness is the trigger of perlocutionary content, rendering the assertion an 'insult'.

(1) *T'es une femme je serai jamais d'accord avec toi pour du foot*
(You're a woman I'll never agree with you about football)

(2) *les femmes qui sont en plus Dijonnaise ne parlez pas de foot sivouplai c'est comme si un aveugle manchot parler de passer le permis*
(women who are also from Dijon please don't talk about football it's as if a one-handed blind person was thinking about getting a driving license)

Descriptive assertions are not directed to an addressee: the target can be a woman, a group of women, or all women, it can be named but is not the addressee. Descriptive assertions are in the third person and thus may have a lower impact on the receiver in comparison with second person assertions. They do not commit the addressee to the truth of the content by soliciting a response. They report generic content (Mari et al., 2012). Linguistic clues can be the presence of a named entity as the target or use of generalizing terms, as shown in (3) and (4).

(3) *Anne Hidalgo est une femme. Les femmes aiment faire le ménage. Anne Hidalgo devrait donc nettoyer elle-même les rues de Paris*
(Anne Hidalgo is a woman. Women love cleaning the house. Anne Hidalgo should clean the streets of Paris herself)

(4) *une femme a besoin d'amour de remplir son frigo, si l'homme peut le lui apporter en contrepartie de ses services (ménages, cuisine, etc) j'vois pas elle aurait besoin de quoi d'autre*
(A woman needs love, to fill the fridge, if a man can give this to her in return for her services (housework, cooking, etc), I don't see whatelse she needs)

Finally, in **reported assertions**, the sexist content is a report of an experience or a denunciation of a sexist behaviour. They may elicit an even lower commitment on behalf of the addressee. The speaker is not committed to the truth of a reported content (as in *I heard that you were coming*

[5]The translations might not feel natural. Indeed, we kept the same words in English as in French in order to better illustrate the type/semantic of words that are used, keeping in mind that tweets are often not well-written in French as well as in English.

too). However, when reporting sexist content, the speaker is still conveying a lack of commitment, and a general sense of disapproval or dismissal may emerge. In these messages, we observe the presence of reporting verbs, quotation, locations (as reports often mention public spaces where the experience happened) or specific hashtags, as shown in (5), (6) and (7).

(5) *je m'assoupis dans le métro, je rouvre les yeux en sentant quelque chose de bizarre : la main de l'homme assis à côté de moi sur ma cuisse. #balancetonporc*
(I doze in the subway, I open my eyes feeling something weird: the hand of the man sat next to me on my leg #SquealOnYourPig)

(6) *Mon patron m'a demandé : "qui va cuisiner pour ton mari quand tu seras pas là ?"*
(My boss asked me: "who's going to cook for your husband when you're away?")

(7) *Je ne suis pas une grande fan de Theresa May mais pourquoi parler de "ses escarpins et ses cuissardes vernies" et la traiter d'allumeuse ? #vincenthervouet #sexisme http://eur1.fr/nADYIMw*
(I am not a fan of Theresa May but why talking about "her shoes and varnished boots" and call her a tease? #vincenthervouet #sexism)

As it appears, the three types of assertions have a sexist content, but only the first two ones are really sexist. Indeed, direct and descriptive assertions are first-hand information, whereas reported ones are second-hand information. As such, they may trigger a different reaction from the receiver: in the first two cases, a female receiver can be immediately involved as the target of the sexist dismissal; in the third case, she is the witness of a sexist report.

4 Data and Annotation

Our corpus is new and contains French tweets collected between October 2017 and May 2018. In order to collect sexist and non sexist tweets, we followed Anzovino et al. (2018) approach using: (i) a set of representative keywords: *femme, fille (woman, girl)*, *enceinte (pregnant)*, some activities (*cuisine (cooking), football, ...*), insults, etc., (ii) the names of women/men potentially victims or guilty of sexism (mainly politicians), (iii) specific hashtags to collect stories of sexism experiences[6]: *#balancetonporc, #sexisme, #sexiste, #SexismeOrdinaire, #EnsembleContreLeSexisme, #payetashnek, #payetontaf,* etc. The tweets collected with these hashtags may contain reported sexist acts towards both men and women. Thus, we collected around $205,000$ tweets, among which about $70,000$ contain the specific hashtags.

Given a tweet, annotation consists in assigning it one of the following five categories: direct, descriptive, reporting (as defined in the previous section), non-sexist and no decision. A tweet is **non sexist** when it has no sexist content (it may contain a specific hashtag, but the content is not sexist), as in (8). **No decision** refers to cases where the tweet lacks context, or when the sexist content is not in the text but only in a photo, video, or URL (because we cannot process them).

(8) *La créatrice du #balancetonporc attaquée en justice pour diffamation*
(France's #MeToo creator on trial for defamation)

300 tweets have been used for the training of 5 annotators (they are master's degree students (3 female and 2 male) in Communication and Gender) and then removed from the corpus. Then, 1,000 tweets have been annotated by all annotators so that the inter-annotator agreement could be computed. Although the perception of sexism is often considered as subjective, the average Cohen's Kappa is 0.72 for sexist content/non-sexist/no decision categories and 0.71 for direct/descriptive/reporting/non-sexist/no decision categories which means a strong agreement. We noticed that the kappa scores between female annotators are very close to the one between male annotators. For these 1,000 tweets, the final labels have been assigned according to a majority vote.

Finally, a total of $11,834$ tweets have been annotated according to the guidelines after removing 1,053 tweets annotated as "no decision". Among them, 65.80% are non-sexist and 34.20% with sexist content (79.61% reporting, 1.12% are direct and 19.27% descriptive). We then divided the corpus into train and test sets[7] (cf. Table 1).

5 Identifying Reports of Sexist Acts

To identify reported assertions, we performed three classification tasks: (BIN) sexist content vs. non-

[6]The distribution of these hashtags is very similar in both non sexist and sexist tweets which reduces considerably the bias while collecting the data.

[7]All the hyperparameters were tuned on the validation set (20% of the training dataset), such that the best validation error was produced.

	Sexist content 4,047		Non sexist 7,787
Train	**direct+descriptive**	**reporting**	
	38 + 599 (= 637)	2,559	6,255
Test	**direct+descriptive**	**reporting**	
	7 + 181 (= 188)	663	1,532

Table 1: Tweet distribution in train/test datasets.

sexist, (3-CLASS) sexist tweets (i.e., direct and descriptive) vs. reporting tweets vs. non-sexist; and (CASC) a cascade classification with sexist content vs. non-sexist in the first stage, followed by reporting vs. non-reporting in the second stage. To this end, we experiment with several deep learning models[8] including best performing state of the art models for sexism detection.

CNN. This model has already been used in Karlekar and Bansal (2018). It uses pre-trained on Wikipedia and Common Crawl FastText French word vectors and three 1D Convolutional layers, each one using 100 filters and a stride of 1, but different window sizes (2, 3, and 4 respectively) with a ReLU activation function. We further downsample the output of these layers by a 1D max pooling layer (with a pool size of 4), and we feed its output to the final softmax layer.

CNN-LSTM. This model is similar to Karlekar and Bansal (2018) and (Parikh et al., 2019) except that we used word-level embeddings instead of character/sentence-level as the results were lower. It is based on the previous CNN model by adding an LSTM layer[9] (capable of capturing the order of a sequence) that takes its input from the max pooling layer. Next, a global max pooling layer feeds the highest value in each timestep dimension to a final softmax layer.

BiLSTM with attention. This model, also used by (Parikh et al., 2019), relies on a Bidirectional LSTM with an attention mechanism that attends over all hidden states and generates attention coefficients. The hidden states were then averaged using the attention coefficients in order to generate the final state, which was then fed to a one-layer feed-forward network in order to obtain the final label prediction. We experimented with different hidden state vector sizes, dropout values and attention vector sizes. The results reported in this paper were obtained by using 300 hidden units, an 150 attention vector, a dropout of 50% and the Adam optimizer with a learning rate of 10^{-3}.

BERT$_{\text{base}}$. It uses the pre-trained BERT model (BERT-Base, Multilingual Cased) (Devlin et al., 2019) on top of which we added an untrained layer of neurons. We then used the HuggingFace's PyTorch implementation of BERT (Wolf et al., 2019) that we trained for 3 epochs.

BERT^R. We observed that about 47% of the tweets embed at least one URL. Due to the short length of a tweet, this is useful for amplifying the message, while also minimizing the time it takes to compose it. In order to feed more information to the classifier, instead of removing or replacing the URLs with replacement tokens as usually done in hate speech detection, we propose to substitute them with the title found at the given URL[10]. In addition, and based on the assumption that word embeddings capture the meaning of words better than emoji embeddings capture the meaning of emojis, we followed the strategy proposed by (Singh et al., 2019) and replaced all the emojis with their detailed descriptions[11]. Replacing URLs and emojis improved the results for all the models we have tested, so we give here only the results obtained after these replacements.

BERT$^R_{\text{own_emb + base}}$. Following (Parikh et al., 2019), we also experiment stacking multiple embeddings. We tailored a pre-trained BERT model[12] for which we used the whole non annotated dataset (i.e., 205,000 tweets). The original BERT model uses a WordPiece tokenizer, which is not available in OpenSource. Instead, we used a SentencePiece[13] tokenizer in unigram mode. Training the model using the Google Cloud infrastructure with the default parameters for 1 million steps took approximately 3 days.

BERT$^R_{\text{features}}$. We relied on state of the art features that have shown to be useful for the task of hate speech detection: *Surface features* (tweet length in words, the presence of personal

[8] We also experiment with standard feature-based models, but the results were lower.

[9] We also experimented with GRU following (Zhang and Luo, 2018), but the results were not conclusive.

[10] In case a particular web page is not available anymore, the URL is removed from the tweet.

[11] We relied on a manually built emoji lexicon that contains 1,644 emojis along with their polarity and detailed description.

[12] We experimented with different configurations by incorporating different French pre-trained embeddings available: Glove (Pennington et al., 2014), FastText (Grave et al., 2018), Flair (Akbik et al., 2018) and CamemBERT (Martin et al., 2019) but none of the configurations were able to achieve results better than BERT$_{\text{base}}$.

[13] https://github.com/google/sentencepiece

pronoun and third-person pronoun, punctuation marks, URLs, images, hashtags, @userMentions and the number of words written in capital), *Emoji features*[11] (number of positive and negative emojis), *Opinion features* (number of positive, negative and neutral words in each tweet relying on opinion (Benamara et al., 2014), emotion (Piolat and Bannour, 2009) and slang French lexicons. We also account for hedges (negation and modality), reporting verbs, imperative verbs, and verbs used for giving advice.

BERT^R$_{\text{gen}}$. Sexism is often expressed by using gender stereotypes, i.e., ideas whereby women and men are arbitrarily assigned characteristics and roles determined and limited by their gender. In order to force the classifier to learn from generalized concept rather than words which may be rare in the corpus, we adopt several replacement combinations extending (Badjatiya et al., 2017)'s approach consisting in replacing some words/expressions that trigger sexist content by their generalized term. However, instead of using a flat list composed of most frequent words that appear in a particular class and then replace them by similarity relationships, we rather rely on manually built lists of words[14] often used in sexist language (hereafter `<SexistVocabulary>`): **designations** (around 10 words such as *femme (woman), fille (girl), nana (doll), ...*), **insults** (around 400 words/expressions extracted from GLAWI (Hathout and Sajous, 2016), a machine-readable French Dictionary); and 130 gender stereotyped words grouped according to the following taxonomy as usually defined in gender studies (see Section 2): **physical characteristics** (e.g. *petite (little), bouche (mouth), robe (dress), ...* for women; *petit (little), gros (fat), ...* for men), **behavioural characteristics** (e.g. *bavarde (gossipy), jalouse (jealous), tendre (loving), ...* for women; *macho, viril (virile), ...* for men), and **type of activities** (e.g. *mère (mother), cuisine (cooking), infirmière (nurse), ...* for women; *football, médecin (doctor), ...* for men). Only 1% of all these words have been used as keywords to collect the corpus.

In addition, we also built two other lists: **names** (952/832 female/male firstnames to detect named entities) and around 170 words/expressions for **places** as they are mainly useful for detection of reporting messages since they represent public spaces where sexist acts may occur.(e.g. *métro (subway), rue (street), bureau (office), ...*).

We experimented with distinct generalization strategies: hypernym replacement **gen(Hypernym)** (e.g., *little* is replaced by `<PhysicalCharacteristics>`), gendered hypernym replacement **gen(Hypernym_gendered)** (e.g., *dress* is replaced by `<femalePhysicalCharacteristics>`) as well as generic replacement **gen(SexistVocabulary)** (e.g., both *little* and *doll* are replaced by the same tag `<SexistVocabulary>`), etc., where X in **BERT^R$_{\text{features+X}}$** indicates the adopted replacement strategy.

6 Results

6.1 BIN and 3-CLASS results

Table 2 presents the results for the best state of the art models for the task of sexism detection (CNN, BiLSTM with attention, CNN-LSTM) applied on the BIN task in terms of accuracy (A), macro-averaged F-score (F), precision (P) and recall (R) with the best results in bold. None of these models were able to achieve results better than BERT$_{\text{base}}$. For this reason, we chose BERT$_{\text{base}}$ as our baseline and trained it on top of several vectorial representations, as explained in Section 5.

CLASSIFIER	A	F	P	R
CNN	0.684	0.601	0.635	0.571
CNN+LSTM	0.676	**0.640**	0.623	0.657
BiLSTM$_{\text{attention}}$	**0.695**	0.527	0.501	0.554
BERT$_{\text{base}}$	**0.773**	**0.723**	**0.726**	**0.721**

Table 2: Results for BIN classification.

As shown in Table 3, we observe that training BERT with stacked embeddings did not improve over BERT$_{\text{base}}$. Replacing URLs and emojis with respectively the words within the title link and emoji description boosts the results by 1.7% and 1.2% in terms of accuracy while adding linguistic features to the embeddings increases the results for both the BIN and 3-CLASS configurations. We, therefore, keep BERT^R$_{\text{features}}$ as basis for the rest of the models. Concerning the generalization strategies, all replacements were productive and outperformed all the previous models, observing that gendered replacements are better. This shows that forcing the classifier to learn from general concepts is a good strategy for sexism content detection. In particular, we observe that the best replacement depends on the task: For BIN, it is place and gen-

[14]Following (Badjatiya et al., 2017), we also experiment with automatic word lists but the results were not conclusive as frequent words were too generic and not representative of the problem we want to solve.

CLASSIFIER	BIN				3-CLASS			
	A	F	P	R	A	F	P	R
$BERT_{base}$	0.773	0.723	0.726	0.721	0.714	0.540	0.572	0.515
$BERT^{R}$	0.790	0.762	0.767	0.759	0.726	0.567	0.609	0.531
$BERT^{R}_{\text{own\_emb + base}}$	0.768	0.751	0.712	0.795	0.708	0.526	0.605	0.513
$BERT^{R}_{\text{features}}$	0.795	0.787	0.819	0.761	0.754	0.588	0.625	0.556
$BERT^{R}_{\text{features + gen(Hypernym)}}$	0.806	0.804	0.835	0.776	0.763	0.614	0.649	0.598
$BERT^{R}_{\text{features + gen(Hypernym\_gendered)}}$	0.809	0.807	0.840	0.777	0.767	0.635	0.663	0.620
$BERT^{R}_{\text{features + gen(Name)}}$	0.790	0.796	0.830	0.766	0.755	0.620	0.656	0.606
$BERT^{R}_{\text{features + gen(Name\_gendered)}}$	0.815	0.806	0.841	0.775	0.760	0.643	0.665	0.630
$BERT^{R}_{\text{features + gen(SexistVocabulary\_gendered)}}$	0.801	0.807	0.836	0.781	0.764	0.635	0.654	0.627
$BERT^{R}_{\text{features + gen(Place)}}$	0.826	0.813	0.848	0.782	0.769	**0.655**	0.673	**0.646**
$BERT^{R}_{\text{features + gen(Place + Hypernym)}}$	0.803	0.799	0.836	0.766	0.758	0.622	0.654	0.610
$BERT^{R}_{\text{features + gen(Place + Hypernym\_gendered)}}$	0.819	0.811	0.846	0.779	**0.771**	0.652	**0.689**	0.630
$BERT^{R}_{\text{features + gen(Place + Name\_gendered)}}$	**0.837**	**0.824**	**0.865**	**0.787**	0.769	0.629	0.657	0.615
$BERT^{R}_{\text{features + gen(Place+Hypernym\_gendered+Name\_gendered)}}$	0.819	0.818	0.857	0.783	0.764	0.634	0.662	0.618

Table 3: Results for most productive models for BIN and 3-CLASS classification.

dered names whereas for 3-CLASS it is place and gendered hypernym. In both cases, replacing only public spaces with the generic `<location>` was one of the best strategy with 0.826 and 0.769 accuracy for respectively BIN and 3-CLASS. Multiple replacements (cf. last line in the table) were however, less productive.

Table 4 further details the results per class for the best performing systems for each task (i.e., those in bold in Table 3). For the 3-CLASS, we observe that the results are lower for the sexist content (direct and descriptive) class, but this might also be a consequence of the low number of instances annotated as such[15].

Task	Class	F	P	R
BIN	non sexist	0.874	0.894	0.855
	sexist	0.773	0.836	0.719
	overall	0.824	0.865	0.787
3-CLASS	non sexist	0.849	0.855	0.842
	reporting	0.666	0.633	0.703
	sexist	0.452	0.532	0.392
	overall	0.655	0.673	0.646
CASC	non sexist	**0.882**	0.912	0.855
	reporting	**0.942**	0.919	0.975
	sexist	**0.791**	0.768	0.816
	A = 0.831			
	overall	0.717	0.724	0.709

Table 4: Results per class for the three tasks.

6.2 CASC results

Cascading models are known for being very accurate and can be used in the context of moderation as we cannot afford to take actions against users that are following the guidelines and policies. In the first stage we used the best performing model for sexist content vs. non sexist classification (i.e., $BERT^{R}_{\text{gen(Place+Name\_gendered)}}$). The instances classified as containing a sexist content by the first model were further used as the testing set for the second model (the best performing model for the 3-CLASS classification task in terms of F-score, i.e., $BERT^{R}_{\text{gen(Place)}}$). In Table 4, the results corresponding to the non-sexist class of CASC classifier present the improvement brought by the second stage classifier, i.e., it was able to correct (predict as non-sexist) instances that were misclassified during the first stage. The last line of Table 4 presents the overall results obtained after the two stages of classification. The results show an improvement over the best system of 3-CLASS, proving the usefulness of a cascading approach with an increasing system complexity.

6.3 Discussion

A manual error analysis shows that misclassification cases are due to several factors, among which humor and satire (as in (9)) or the use of stereotypes (as in (10)), mainly because they are not expressed by a single word or expression but by metaphors. In the examples below, the underlined words highlight the leading cause of misclassification.

(9) *Ma femme est hystorique. C'est comme hystérique, sauf que lorsqu'elle pète un câble elle me sort des vieux dossiers.*
(*My wife is hystorical. That's like hysterical, except that when she's angry she pulls out old files*)

[15] We tried augmenting the number of instances in these classes by replacing the words/phrases that belong to the sexist vocabulary and stereotyped words list (cf. Section5) with the top 10 word2vec neighbours (i.e., for each instance we obtain 10 more) but the results were not conclusive. More accurate data augmentation techniques can be investigated.

(10) *je demande pas ce qu'elle a fait sous le bureau pour arriver à se plateau*
(*I'm not asking what she did under the desk to be on this TV set*)

In particular for reporting tweets, we found many misclassified messages without any reporting verb or quotes as in (11), but also messages denunciating sexism using situational irony as in (12).

(11) *Royal les rendrait elle tous fous? Alain Destrem (UMP): Ségolène Royal en boubou bleu, ça me rappelle ma femme de ménage !*
(*Does Royal make them all crazy? Alain Destrem (UMP): Ségolène Royal wearing a blue boubou, it reminds me my cleaning woman!*)

(12) *Continuons à communier... Notre héros national avait des comptes en Suisse et n'était pas loin du #balancetonporc... Mais bon communions, rassemblons nous...*
(*Let's keep on be united... Our national hero had bank accounts in Switzerland and was not far from #SquealOnYourPig... But OK let's be united, let's get together...*)

7 Conclusion

In this paper, we have presented the first approach to detect reports/denunciations of sexism from real sexist content that are directly addressed to a target or describes a target. We proposed a new dataset of about $12,000$ French tweets annotated according to a new characterization of sexist content inspired from both speech act theory and discourse studies in gender. We then experimented with several deep learning models in binary, three classes and a cascade classifier configurations, showing that BERT trained on word embeddings, linguistic features and generalization strategies (i.e., place and hypernym replacements) achieved the best results for all the configurations, and that cascade classification allows to successfully correct misclassified non-sexist messages. These results are encouraging and demonstrate that detecting reporting assertions of sexism is possible, which is a first step towards automatic offensive content moderation. In the future, we plan to develop more complex models to be added in the next stages of the cascade classifier as well as automatically identify irony, gender stereotypes and sexist vocabulary.

Acknowledgments

This work is funded by the Institut Carnot Cognition under the project SESAME.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018*, pages 57–64.

John L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-based Generalizations. In *The World Wide Web Conference*, pages 49–59.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Azy Barak. 2005. Sexual harassment on the Internet. *Social Science Computer Review*, 23(1).

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Farah Benamara, Véronique Moriceau, and Yvette Yannick Mathieu. 2014. Fine-grained semantic categorization of opinion expressions for consensus detection (in French). In *DEFT 2014 Workshop: Text Mining Challenge*, pages 36–44.

Claudia Bianchi. 2014. The speech acts account of derogatory epithets: some critical notes. *Liber Amicorum Pascal Engel*.

Laurence Biscarrat, Marlène Coulomb-Gully, and Cécile Méadel. 2016. One is not born a female CEO and...won't become one! *Gender equality and the media - a challenge for Europe. Routledge, ECREA Book Series*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364.

Simone Bonnafous. 2003. "Femme politique" : une question de genre ? *Réseaux*, 120.

Danielle Bousquet, Françoise Vouillot, Margaux Collet, and Marion Oderda. 2019. 1er état des lieux du sexisme en France. Technical report, Haut Conseil à l'Egalité entre les femmes et les hommes. http://www.haut-conseil-egalite.gouv.fr/IMG/pdf/hce_etatdeslieux-sexisme-vf-2.pdf.

Deborah Cameron. 1992. *Feminism and Linguistic Theory.* Palgrave Macmillan.

Bianca Cepollaro. 2015. In defence of a presuppositional account of slurs. *Language Sciences*, 52.

Pat K Chew and Lauren K Kelley-Chew. 2007. Subtly sexist language. *Colum. J. Gender & L.*, 16.

European Commission COM. 2017. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Tackling illegal content online, towards an enhanced responsibility of online platforms. Technical report, COM (2017), 555 final. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017DC0555.

Marlène Coulomb-Gully. 2012. *Présidente : le grand défi - femme, politique et medias*. Paris, Payot/Éd. Rivages.

Haoyun Dai and Xiaodong Xu. 2014. Sexism in News: A Comparative Study on the Portray of Female and Male Politicians in The New York Times. *Open Journal of Modern Linguistics*, 4.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International Conference on Web and Social Media*, pages 512–515.

Richard Delgado, Adrien Katherine Wing, and Jean Stefancic. 2015. Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling. In *Law Unbound!*, pages 223–228. Routledge.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating Bias in Word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, pages 879–887.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding. 2018. Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018*, pages 42–51.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). volume 2150 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Anastasia Giannakidou and Alda Mari. 2021. *(Non) Veridicality in grammar and thought. Mood, Modality and Propositional Attitudes*. The University of Chicago Press.

Peter Glick and Susan T. Fiske. 1996. The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology*, 70(3).

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 609–614.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of LREC*.

Nabil Hathout and Franck Sajous. 2016. Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1369–1376.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16.

Sweta Karlekar and Mohit Bansal. 2018. SafeCity: Understanding Diverse Forms of Sexual Harassment Personal Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811.

Rae Langton. 2012. Beyond Belief: Pragmatics in Hate Speech and Pornography. *Speech and Harm: Controversies Over Free Speech*.

Michelle M Lazar. 2007. Feminist critical discourse analysis: Articulating a feminist discourse praxis. *Critical Discourse Studies*, 4(2).

Kate Manne. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Alda Mari, Claire Beyssade, and Fabio Del Prete. 2012. *Genericity*, volume 43. Oxford University Press.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. *arXiv e-prints*, page arXiv:1911.03894.

Mari J Matsuda, Charles R Lawrence III, Richard Delgado, and Kimberle Williams Crenshaw. 1993. Words that wound: Critical race theory. *Assaultive Speech, and the First Amendment*, 5.

Jessica Megarry. 2014. Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47.

Sara Mills. 2008. *Language and sexism*. Cambridge University Press Cambridge.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of EMNLP-IJCNLP*.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label Categorization of Accounts of Sexism using a Neural Framework. In *Proceedings of EMNLP*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Annie Piolat and Rachid Bannour. 2009. An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing. *Current psychology letters*, 25.

Paul Portner. 2018. *Mood*. Oxford University Press.

Christopher Potts. 2005. *The logic of conventional implicatures*. Oxford Studies in Theoretical Linguistics.

Hemant Purohit, Tanvi Banerjee, Andrew Hampton, Valerie L. Shalin, Nayanesh Bhandutia, and Amit P. Sheth. 2016. Gender-Based Violence in 140 characters or Fewer: A #BigData Case Study of Twitter. *First Monday*, 21(1).

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An Analysis of Domestic Abuse Discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs. In *Proceedings of the Second Workshop on Abusive Language Online*.

Abhishek Singh, Eduardo Blanco, and Wei Jin. 2019. Incorporating Emoji Descriptions Improves Tweet Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2096–2101.

Jeremy Waldron. 2012. *The harm in hate speech*. Harvard University Press.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Ziqi Zhang and Lei Luo. 2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *arXiv preprint arXiv:1803.03662*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.