# An Annotated Corpus for Sexism Detection in French Tweets

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, Marlène Coulomb-Gully

*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1397–1403
Marseille, 11–16 May 2020


# An Annotated Corpus for Sexism Detection in French Tweets

**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

**Abstract**

Social media networks have become a space where users are free to relate their opinions and sentiments which may lead to a large spreading of hatred or abusive messages which have to be moderated. This paper presents the first French corpus annotated for sexism detection composed of about 12,000 tweets. In a context of offensive content mediation on social media now regulated by European laws, we think that it is important to be able to detect automatically not only sexist content but also to identify if a message with a sexist content is really sexist (i.e. addressed to a woman or describing a woman or women in general) or is a story of sexism experienced by a woman. This point is the novelty of our annotation scheme. We also propose some preliminary results for sexism detection obtained with a deep learning approach. Our experiments show encouraging results.

**Keywords:** sexism detection, social media, corpus, speech acts

## 1. Introduction

Social media networks such as Facebook, Twitter, blogs and forums, have become a space where users are free to relate events, personal experiences, but also opinions and sentiments about products, events or other people. This may lead to a large spreading of hatred or abusive messages which have to be moderated. In particular, theses messages may express threats, harassment, intimidation or "disparage a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics" (Nockleby, 2000). Although some countries, such as the United States, where hate speech is protected under the First Amendment as freedom of expression (Massaro, 1990), many other countries, such as France, have laws prohibiting it, laws that extend to the internet and social media. For instance, since the French law of 27 January 2017 related to equality and citizenship, penalties due to discrimination are doubled and sexism is now considered as an aggravating factor.

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 and it mainly affects women and girls. It can take several forms: sexist remarks, gestures, behaviours, practices, from insults to rape or murder. As mentioned in *Combating Sexist Hate Speech*, a report of the Council of Europe[1], "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". Its psychological, emotional and/or physical impacts can be severe.

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).

These last years, social media and web platforms have offered a large space to sexist hate speech (in France, a recent study of the High Council to Equality reports that 10% of sexist abuses come from social media[2]) but also allow to share stories of sexism experienced by women (see "The Everyday Sexism Project"[3] available in many languages, "Paye ta shnek"[4] in French, or hashtags such as #metoo or #balancetonporc). In this context, it is important to automatically detect sexist messages on social platforms and possibly to prevent the widespreading of gender stereotypes, especially towards young people.

In this paper, we propose:

1. A novel characterization of sexist content inspired by speech acts theory (Austin, 1962). We distinguish different types of sexist content according to their perlocutionary force: sexist hate speech *directly addressed* to a target, *descriptive assertions* of a woman or women in general, or *reported assertions* that relate a story of sexism experienced by a woman.

2. The first French dataset of about 12,000 tweets annotated for sexism detection according to this new characterization that is freely available for the research community[5].

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

[2] http://www.haut-conseil-egalite.gouv.fr/IMG/pdf/hce_etatdeslieux-sexisme-vf-2.pdf

[3] https://everydaysexism.com/

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

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

3. A set of experiments to detect sexist content in messages relying both on standard feature-based and deep learning approaches using either contextualized and non contextualized embeddings. Our results are encouraging and constitute the novel state of the art on sexism detection in French.

The paper is organized as follows. Section 2. presents state of the art. Section 3. describes our data, the characterization of sexism content we propose and the annotation scheme. Section 4. presents the experiments we carried out on our data. We conclude providing some perspectives for future work.

## 2. Related Work

In corpus construction, sexism is often considered as "hate speech" (Golbeck et al., 2017) and has been widely studied as such in a purpose of automatic detection (Badjatiya et al., 2017), (Waseem and Hovy, 2016). Hate speech detection covers mainly the detection of explicit racist, abusive or offensive textual content using supervised machine learning approaches either with bag-of-words, or dedicated lexicons, word embeddings, clustering, or author profile (see (Schmidt and Wiegand, 2017) for a survey).

Current work consider sexism detection (sexist vs. non sexist) or sexism classification (identifying the type of sexist behaviours). In the last case, categories are most often mutually exclusive (e.g., harassment, threat, physical violence, body shaming, benevolent, etc.) (Jha and Mamidi, 2017; Sharifirad et al., 2018), except in (Parikh et al., 2019) who consider messages of sexism experienced by women in the "Everyday Sexism Project" web site and whose categories are not mutually exclusive.

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. The datasets were composed of about 4.000 tweets for each language. Most participants used Support Vector Machines (SVM) and ensemble of classifiers for both tasks with features such as n-grams and opinions (Fersini et al., 2018). Very few participants used deep learning approaches with word embeddings and best results were obtained with SVM models. These datasets have also been used in the *Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter* shared task at SemEval 2019. The tasks were the same as those of AMI except that it concerned not only sexism against women but also hate speech against immigrants. The best results were obtained with a SVM model using sentence embeddings as feature (Indurthi et al., 2019).

As far as we know, no work have addressed sexism detection in French. Furthermore, in a context of offensive content moderation on social media (see the recommendations of the European commission[6]), we think 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 and reports/denunciations of sexism experiences. Indeed, whereas messages could be reported and moderated in the first case, messages reporting sexism experiences should not be moderated.

## 3. Data and Annotation

### 3.1. Data Collection

Our corpus is new and extends the one we used in (Chiril et al., 2019). It 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:

- a set of representative keywords: *femme, fille (woman, girl)*, *enceinte (pregnant)*, some activities (*cuisine (cooking), football, ...*), insults, etc.,
- the names of women/men potentially victims or guilty of sexism (mainly politicians): *Ségolène Royal, Nadine Morano, Theresa May, Hillary Clinton, Dominique Strauss-Kahn, Nicolas Hulot*, etc.,
- specific hashtags to collect stories of sexism experiences: *#balancetonporc, #sexisme, #sexiste, #SexismeOrdinaire, #EnsembleContreLeSexisme, #payetashnek, #payetontaf,* etc..

Thus, we collected around 115,000 tweets among which about 30,000 contain the specific hashtags. Before detailing our annotation scheme and the result of the annotation procedure, the next subsection presents the theoretical backgrounds on which we based our study.

### 3.2. 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). Our study pursues a different line of analysis, whereby speech acts bearing on derogatory content are ranked according to their perlocutionary force and assertions are classified as more or less direct.

Specifically, in order to make emerge different degrees of downgrading tones, 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 (ElSherief et al., 2018; Ousidhoum et al., 2019). ElSherief et al. (2018) consider that directed hate speech is explicitly directed at a

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017DC0555

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, we newly consider three different stages in the scale of 'directedness' of an assertion: assertions directed to the addressee, descriptive assertions and reported assertions.

Sexist content in **directed assertions** is explicitly directed at a woman but contrary to both approaches cited above, it can also be directed at a group of women or all women. Across the different classifications of speech acts (Portner, 2009) there is a basic distinction that cuts across different types of speech acts: directedness and indirectedness. Questions and imperatives are 'direct' in the sense that they require that the addressee performs an action (responding (with questions) or acting (with imperatives)). Assertions themselves can be direct or indirect. They are direct when they are in the second person ('you').

**Descriptive assertions** are not directed to the 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 addressee in comparison with second person assertions. They report generic content.

Finally, **reported assertions** may elicit an even lower commitment on behalf of the addressee (see (Portner, 2009; Giannakidou and Mari, to appear) for a general discussion on evidentiality and reportativity). The speaker is not committed to the truth of a reported content (as in *I heard that you were coming too*). However, when reporting sexist content, the speaker is still conveying lack of commitment, and a general sense of disapproval or dismissal may emerge.

As it appears, the first two types of assertions are first-hand information, whereas the third type is second hand information. As such, they may trigger a different reaction from the addressee: in the first two cases the addressee is immediately involved as the target of the sexist dismissal; in the third case, she is the witness of a sexist report.

### 3.3. Annotation Guidelines

We used a set of 150 tweets to define the annotation guidelines. As already said before, the novelty of our approach is that we want to identify not only sexist content in tweets but also if the tweet is really sexist (i.e. directly address to a target or describing a target) or is a story of sexism experienced by a woman. Given a tweet, annotation consists in assigning it one of the following three categories:

**(i) Sexist content**: it can be either **direct** (cf. (1) and (2)), **descriptive** (cf. (3) and (3)) or **reporting** (cf. (5) and (6)). The first two are real sexist messages but not the last one as reporting tweets must not be considered as sexist in a context of moderation.

**Direct sexist content**, directly addressed to a woman or a group of women, generally uses second person pronoun/verb and imperatives, as shown in the examples below (linguistic clues are underlined).

(1) *t'es une femme pq tu veux parler de foot?*
*(You're a woman why do you want to talk 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)*

In **descriptive sexist content** where the tweet describes a woman or women in general, clues can be presence of a named entity or use of generalizing terms.

(3) *Theresa May succède à David Cameron. Pas mieux qu'une femme pour faire le ménage.*
*(Theresa May succeeds David Cameron. No better at cleaning than a woman.)*

(4) *La place de la femme dans la societe moderne, est tres precise elle est dans la cuisine.*
*(The place of a woman in modern society is clear, it is in the kitchen.)*

When the sexist content is in fact a **report** of a sexism experience or a denunciation of a sexist behaviour, we observe the presence of reporting verbs, quotation and specific hashtags.

(5) *il m'a dit "normal qu'elles soient moins payer pcq enceinte=moins de temps au travail"*
*(he told me "normal that they are paid less because pregnant=less working time")*

(6) *#RolandGarros : une petite remarque bien #sexiste sur l'émotivité des joueuses de tennis dans Stade2*
*(#RolandGarros:a little #sexist comment about the emotionality of female tennis players in Stade2 TV Show)*

**(ii) Non sexist**: when the tweet has no sexist content (it may contain a specific hashtag but the content is not sexist), as in (7) and (8).

(7) *Paris Match : journal d'investigation*
*(Paris Match: an investigative journal)*

(8) *Laetitia Casta pas d'accord avec #balancetonporc*
*(Laetitia Casta disagrees with #SquealOnYourPig)*

**(iii) No decision**: when the tweet is not understandable (not well-written, lack of context) or when the sexist content is not in the text but only in a photo, video, or URL (because we cannot process them), as in (9).

(9) *J'ai envie de poster ça sur facebook mais j'ai peur des commentaires ... pic.twitter.com/kMq(...)*
*(I'd like to post this on Facebook but I fear comments... pic.twitter.com/kMq(...))*

### 3.4. Manual Annotation

300 tweets have been used for the training of 5 annotators (they are master degree's 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 and 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. 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 the tweets annotated as "no decision". Table 1 shows the distribution of the annotated corpus.

| **Sexist content** | | | **Non sexist** | **Total** |
|---|---|---|---|---|
| 4,047 | | | 7,787 | 11,834 |
| **direct** | **descriptive** | **reporting** | | |
| 45 | 780 | 3,222 | | |

Table 1: Tweet distribution in our French dataset.

## 4. Sexist Content Detection: Preliminary Experiments

Automatically labelling tweets as sexist or not sexist is a challenging task because the language of tweets is full of grammatically and/or syntactic errors and by employing techniques such as sarcasm, satire or irony, the intended nature of the message becomes difficult to detect. For the task at hand, we propose several models ranging from standard bag of words (our baseline) to deep learning models for a sexist content vs. non sexist classification of tweets. To this end, the corpus has been divided into train and test sets. Table 2 shows the distribution of these sets. In the next sections, we detail our models, provide and discuss our results.

| | **Sexist content** | | | **Non sexist** |
|---|---|---|---|---|
| Train | 3,564 | | | 6,255 |
| | **direct** | **descriptive** | **reporting** | |
| | 38 | 599 | 2,559 | |
| Test | 923 | | | 1,532 |
| | **direct** | **descriptive** | **reporting** | |
| | 7 | 181 | 663 | |

Table 2: Tweet distribution in train/test datasets.

### 4.1. Models

**Baseline (SVM BoW).** Due to the noise in the data, we performed standard text pre-processing by removing user mentions, URLs, RT, stop words, degraded stop words and the words containing less than 3 characters. We experimented with several feature-based machine learning algorithms (Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest) in order to evaluate and select the best performing one. Hereby, the baseline is a Support Vector Machine (linear kernel, $C = 0.1$) with unigrams, bigrams and trigrams TF/IDF.

**SVM BoW with URL/emoji replacement.** After inspection, we observed that about 47% of the tweets embed in their text at least one URL. Due to the short length of a tweet, incorporating URLs is useful for amplifying the message, while also minimizing the time it takes to compose the message. By ignoring the content present at a shared URL, an important part of the meaning of the message is lost, as it becomes harder to identify the context. In order to feed more information to the classifier, instead of removing or replacing the URLs with replacement tokens, we propose to substitute them with the title found at the given URL[7].
Emotional content holds an important place in language, as sometimes, what people write may not actually reflect their feelings at the time of writing those words. Emojis have become very popular in social media and are interesting because they encode meaning that otherwise would require more than one word to convey (e.g., grinning face, smiling face with 3 hearts, beaming face with smiling eyes). 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 we replaced all the emojis with their detailed descriptions[8].

After replacing the URLs and emojis as described above, several deep learning models were also trained and evaluated on our dataset. For the following models we used pre-trained word embeddings on Wikipedia and Common Crawl FastText French word vectors with an embedding dimension of 300 (Grave et al., 2018) and we run all the experiments for maximum 100 epochs, with a patience of 10 and batch size of 64[9].

**CNN.** This model uses 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 down sample 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 extends the previous CNN model by adding a LSTM layer[10] (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 uses 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 hid-

[7] In case a particular webpage is not available anymore, the URL is removed from the tweet.

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

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

[10] We also experimented with GRU, but the results were lower.

den 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** (Devlin et al., 2019). For this model we made use of the pre-trained BERT model (BERT-Base, Multilingual Cased) on top of which we added an untrained layer of neurons. For training the new model we used the HuggingFace's PyTorch implementation of BERT (Wolf et al., 2019) that we trained for 3 epochs.

### 4.2. Results

Table 3 shows how the experiments were set up and presents the results in terms of accuracy, macro-averaged F-score, precision and recall.

| Classifier | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|
| SVM BoW | 0.535 | 0.5 | 0.473 | 0.486 |
| SVM BoW + URL/emoji | 0.596 | **0.818** | 0.553 | 0.659 |
| CNN | 0.684 | 0.635 | 0.571 | 0.601 |
| CNN+LSTM | 0.676 | 0.623 | 0.657 | 0.640 |
| BiLSTM with attention | 0.695 | 0.501 | 0.554 | 0.527 |
| BERT | **0.790** | 0.767 | 0.759 | **0.762** |

Table 3: Results for sexist content vs. non sexist classification.

The SVM classifier applied to the dataset having the URLs and emojis replaced by their detailed description improves the results of our baseline, this being the model that also provides the highest precision amid all the models. Among the six models, BERT represents our best performing one in terms of both accuracy and F-score.
Table 4 presents the detailed results for each class (sexist/non sexist content). We note that the results are lower for the sexist content class which leaves enough room for improvement.

| Class | F-score | Precision | Recall | Macro F-score |
|---|---|---|---|---|
| non sexist | 0.843 | 0.832 | 0.856 | 0.762 |
| sexist content | 0.682 | 0.702 | 0.662 | |

Table 4: Results per class with BERT.

### 4.3. Discussion

A manual error analysis of the instances for which our best performing model (BERT) and manual annotation differ shows that in the misclassification of instances that contain a sexist content intervene several factors, among which: the absence of context within the utterance, humor and satire, the use of stereotypes or metaphors. Below, we have provided some examples, the words written in bold highlighting the main cause of misclassification.

**Irony and humor:** In (10), irony should be detected so that the tweet can be classified as "sexist content" (more precisely it is annotated as "reporting" since it denounces ironically those who criticize the way women dress). To this end, we plan to test our approach for irony detection on this corpus (Karoui et al., 2015).

(10) ***Bravo continuez à critiquer*** *Aurore Bergé pour son décolleté et sa jupe courte,* ***vous allez changer le monde comme ça les génies****. Vous avez toujours pas compris qu'on n'a à attendre votre validation pour s'habiller comme on veut ? #SLT*
(**Bravo continue to criticize** Aurore Bergé for her cleavage and her short skirt, **you will change the world like that you geniuses**. You still haven't understood that we are not waiting for your validation to dress how we want? #SLT)

In (11), the tweet contains a wordplay since "balancer" in French means "to denounce" but also "to throw in the trash". Here, the author makes fun of the hashtag; however, there is no sexist content. This tweet has been classified as "sexist content" probably because of the presence of the hashtag.

(11) *#BalanceTonPorc j'ai mis mon cochon à la poubelle mais sa fait rien aider moi*
(#SquealOnYourPig I put my pig in the trash but it doesn't work help me)

**Sexist stereotypes:** (12) has a sexist content using gender stereotypes (women love money and rich men) but not detected as such, probably because the words used for stereotypes are rare and not insults.

(12) *Bientôt 10 ans qu'on est ensemble, j'ai tenté de te larguer mais tu m'obscède tellement que tu es toujours revenue dans ma vie, tu es un peu* ***veinale*** *et* ***coûteuse****, tout le monde dit que tu me fais du mal, mais tu es la, toujours contre ma bouche! #postyourqueen pic.twitter.com/s6OnynFudv*
(It's been 10 years since we've been together, I've tried dumping you, but you're confusing me because you're always coming back into my life, you're **venal** and **costly**, everyone says that you're hurting me, but you are there, always close to my mouth #postyourqueen pic.twitter.com/s6OnynFudv)

(13) also has a sexist content but is misclassified. Here, besides the use of a stereotype (being hysteric is a female stereotype), some reasoning is necessary since the author means that those who use the hashtag #balancetonporc are hysteric.

(13) *quand on participe à une* ***hystérie collective****, il ne faut pas dire n'importe quoi* ***#BalanceTonPorc***
(when you participate in a **collective hysteria**, you shouldn't talk nonsense **#SquealOnYourPig**)

**Need of reasoning:** In the following misclassified example (cf. (14)), the sexist content can be inferred because of the analogy with Google offering food.

(14) *Proposer de congeler ses ovocytes pour éviter que ses employés femmes ne tombent enceinte* ***c'est***

***comme dire que*** *si Google offre de la nourriture dans ces locaux c'est pour qu'ils restent au travail plus longtemps. Et elle travail à la Silicon Valley. #Quotidien*
(Proposing to freeze their oocytes to prevent their female employees from getting pregnant **is like saying that** if Google offers food in its offices it is so that their employees stay at work longer. And she works in Silicon Valley. #QuotidienTVshow)

The sexist content in (15) expressed by a comparison is not recognized but could be easily detected by applying methods for identifying benevolent sexism (expressed by templates such as *aussi bien qu'un homme*) as proposed by (Jha and Mamidi, 2017):

(15) *Leila_Mts Je doute que vous puissiez supporter la cadence infernale du travail de chantier* ***aussi bien qu'un homme*** *le ferait, mais soit.*
(Leila_Mts I doubt that you can handle the hellish rhythm of construction work **as well as a man** would, but so be it.)

## 5. Conclusion

In this paper, we have presented the first corpus of French tweets annotated for sexism detection. It is composed of about 115,000 tweets among which 12,274 are annotated. The novelty of our approach is that not only tweets with a sexist content are labelled but the type of content is also characterized: either the tweet is directly addressed to a target (a woman or all women), describes a target or reports/denounces sexism experienced by a woman. We think that it is important to distinguish between these usages in a context of offensive content moderation on social media since stories of sexism experiences should not be reported. We have experimented several models from standard feature-based to deep learning approaches for a sexist content vs. non sexist classification of tweets. The best results are obtained with BERT. For future work, we plan to add features to our BERT model in order to improve classification based on our error analysis and distinguish tweets labelled as *reporting* from the others.

## 6. Acknowledgements

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

## 7. Bibliographical References

Austin, J. L. (1962). How to Do Things with Words. Oxford University Press.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion.

Bianchi, C. (2014). The speech acts account of derogatory epithets: some critical notes. *Liber Amicorum Pascal Engel*.

Cameron, D. (1992). Feminism and Linguistic Theory. Palgrave Macmillan.

Cepollaro, B. (2015). In defence of a presuppositional account of slurs. *Language Sciences*, 52.

Chiril, P., Benamara, F., Moriceau, V., Coulomb-Gully, M., and Kumar, A. (2019). Multilingual and Multitarget Hate Speech Detection in Tweets. In Proceedings of TALN.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., and Belding, E. (2018). Hate Lingo: A target-based linguistic analysis of hatespeech in social media. In Proceedings of ICWSM.

Fersini, E., Rosso, P., and Anzovino, M. (2018). Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages .

Giannakidou, A. and Mari, A. (to appear). Veridicality in Grammar and Thought: modality, propositional attitudes and negation.

Glick, P. and Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., and Varma, V. (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.

Jha, A. and Mamidi, R. (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.

Karoui, J., Benamara Zitoune, F., Moriceau, V., Aussenac-Gilles, N., and Hadrich Belguith, L. (2015). Towards a contextual pragmatic model to detect irony in tweets. In Proceedings of ACL.

Langton, R. (2012). Beyond Belief: Pragmatics in Hate Speech and Pornography. *Speech and Harm: Controversies Over Free Speech*.

Massaro, T. M. (1990). Equality and freedom of expression: The hate speech dilemma. *Wm. & Mary L. Rev.*, 32:211.

Nockleby, J. T. (2000). Hate speech. In Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In Proceedings of EMNLP-IJCNLP.

Parikh, P., Abburi, H., Badjatiya, P., Krishnan, R., Chhaya, N., Gupta, M., and Varma, V. (2019). Multi-label Categorization of Accounts of Sexism using a Neural Framework. In Proceedings of EMNLP.

Portner, P. (2009). Modality. Oxford University Press.

Potts, C. (2005). The logic of conventional implicatures. Oxford Studies in Theoretical Linguistics.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.

Sharifirad, S., Jafarpour, B., and Matwin, S. (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.

Singh, A., Blanco, E., and Jin, W. (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, Volume 1 (Long and Short Papers), pages 2096–2101, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

## 8. Language Resource References

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic Identification and Classification of Misogynistic Language on Twitter. In International Conference on Applications of Natural Language to Information Systems. Available online at `https://amiibereval2018.wordpress.com/important-dates/data/` (accessed 25 Nov. 2019).

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., Rogers, K. M., Rogers, L., Sarin, M. S., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z., and Wu, D. M. (2017). A Large Labeled Corpus for Online Harassment Research. In Proceedings of the 2017 ACM on Web Science Conference.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL student research workshop. Available online at `https://github.com/ZeerakW/hatespeech` (accessed 25 Nov. 2019).