"Be nice to your wife! The restaurants are closed": Can Gender Stereotype Detection Improve Sexism Classification?

Patricia Chiril and **Farah Benamara** and **Véronique Moriceau**
IRIT, Université de Toulouse, Université Toulouse III - UPS, France
{firstname.lastname}@irit.fr

Abstract

In this paper, we focus on the detection of sexist hate speech against women in tweets studying for the first time the impact of gender stereotype detection on sexism classification. We propose: (1) the first dataset annotated for gender stereotype detection, (2) a new method for data augmentation based on sentence similarity with multilingual external datasets, and (3) a set of deep learning experiments first to detect gender stereotypes and then, to use this auxiliary task for sexism detection. Although the presence of stereotypes does not necessarily entail hateful content, our results show that sexism classification can definitively benefit from gender stereotype detection.

1 Introduction

Stereotypes were originally defined by (Lippmann, 1946) as "pictures in our heads", contending that our imagination is shaped by the pictures we see. This definition explains the way in which opinions are formed and manipulated because of what we trust, that in consequence "leads to stereotypes that are hard to shake". Stereotypes provide information about what a group is like (they are descriptive), but also about why group members are the way they are (they are explanatory).

Although stereotypes can be positive or negative, these generalizations are often linked to negative attitudes towards members of certain social groups (Fiske, 1998). As such, stereotypes represent the root cause of sexism, racism and other inter-group tensions because they convey attributional information that model the way in which stereotyped social group members are being treated by others, as well as the way in which they perceive themselves.

In this paper, we focus on: (1) gender stereotypes (GS hereafter) defined by the Office of the High Commissioner for Human Rights as "a generalised view or preconception about attributes, or characteristics that are or ought to be possessed by women and men or the roles that are or should be performed by men and women", and (2) sexist hate speech which aims according to the Council of Europe 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] In particular, as 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)), it is important to automatically detect sexist messages and possibly to prevent the wide-spreading of GS as they may be used in sexist messages to make generalizations about women, most of the time negative (*e.g., women can't drive*).

GS have been widely studied in psychology, communication studies and social science (Allport et al., 1954; Beike and Sherman, 2014; Crawford et al., 2002; Biscarrat et al., 2016). In NLP, they have been studied mainly to detect or remove gender bias in word embeddings or word association graphs (Bolukbasi et al., 2016; Park et al., 2018; Madaan et al., 2018; Dev and Phillips, 2019; Du et al., 2019) as well as to identify disparity across gender in various applications like co-reference resolution (Zhao et al., 2018), sentiment analysis (Felmlee et al., 2019; Cryan et al., 2020).

In addition to GS, other types of stereotypes have been investigated, such as in the HaSpeeDe 2 shared task (Sanguinetti et al., 2020) which focused on racist stereotypes with tasks for stereotypes and hate speech detection against minority groups. Francesconi et al. (2019) conducted an error analysis on the HaSpeeDe 2018 evaluation campaign (Bosco et al., 2018) concluding that there is a significant correlation between the usage of racist stereotypes and hate speech and that the false positive rate of hateful tweets is slightly higher for tweets that also contain stereotypes. Although sim-

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

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2833–2844
November 7–11, 2021.

ilar correlations have been observed between GS and hate speech from a psychological perspective (García-Sánchez et al., 2019), to our knowledge, no one has empirically measured the impact of GS detection for sexist hate speech classification.

In this paper, we aim to bridge the gap by proposing for the first time an approach for GS detection in tweets as well as a method to inject stereotype information to improve sexism classification. In particular, our contributions are:

(1) The first dataset annotated for GS detection. This dataset contains about 9,200 tweets in French annotated according to different stereotype aspects.[2]

(2) A new method for data augmentation based on sentence similarity with multilingual external resources in order to extend our training dataset (cf. Section 3).

(3) A set of experiments first to detect GS (cf. Section 4) and then, to use this prediction for sexism detection (cf. Section 5). We rely on several deep learning architectures leveraging various sources of linguistic knowledge (label embeddings, generalization strategies based on both manual and automatically generated lexicons) to account for GS and the way sexist contents are expressed in language. Our results show that similarity-based data augmentation is very effective and that sexism classification can definitively benefit from GS detection, beating several strong state of the art baselines for sexist hate speech detection. These results suggest that GS detection is a task by its own that deserves to be studied, for example for educational purpose.

2 Related Work

2.1 Stereotypes in Social Sciences

Stereotypes can be useful for making quick assertions, but the reader should keep in mind that by categorizing people only based on their gender, religion, etc. one has an oversimplified view of reality, which reinforces the perceived boundaries between individuals and seemingly justifies the social implications of role differentiation and social inequality. As gender continues being seen only as a binary categorization, GS not only reflect the differences between women and men, but also impose what men and women should be and how they should behave in regards to different life aspects.

Haines et al. (2016) conducted a study in order to analyze to what extent GS changed over a period of 30 years (in between 1983 and 2014), with participants assessing the likeliness of gendered characteristics (such as traits, behaviours, occupations, physical characteristics) to belong to a typical man or woman. The authors did not find any indication of substantial change of basic stereotypes over time in spite of all the societal changes.

2.2 Stereotype Detection in NLP

Racist stereotypes have been extensively investigated in NLP (Fokkens et al., 2018). For example, the dataset of the HaSpeeDe 2 shared task contains annotated tweets and newspaper headlines, with the main goal of identifying contents that convey hate or prejudice against a given target (immigrants, Muslims and Roma people) with an auxiliary task of determining the presence or absence of a stereotype towards that given target. Among participants, only Lavergne et al. (2020) consider the interaction between hate speech and stereotype detection by employing a multitask learning approach achieving the best scores in the competition. The presence of stereotypes against immigrants has also been annotated in Italian (Sanguinetti et al., 2018) and Spanish political debates (Sánchez-Junquera et al., 2021), the latter being annotated according to a fine-grained taxonomy to capture the positive (threats) and negative dimensions (victims) of stereotypes.

Concerning GS, there are some datasets dedicated to sexist hate speech annotated with stereotype. Among them, Parikh et al. (2019) propose a dataset which contains 13,023 accounts of sexism extracted from the Everyday Sexism Project website manually annotated with 23 labels. The annotation scheme includes two categories for GS: *role stereotyping* (i.e., false generalizations about certain roles being more appropriate for women) and *attribute stereotyping* (i.e., linking women to some physical, psychological, or behavioural qualities). Parikh et al. (2019) classify these messages using LSTM, CNN, CNN-LSTM and BERT models trained on top of several distributional representations (characters, subwords, words and sentences) along with additional linguistic features.

The Automatic Misogyny Identification (AMI) shared task at IberEval and EvalIta 2018 consisted in detecting sexist tweets and then identifying the type of sexist behaviour according to a taxonomy defined by (Anzovino et al., 2018): dis-

[2] https://bit.ly/FrenchGenderStereotypes

credit, 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., 2018).

Besides shared tasks, few studies investigated GS detection. Among them, Felmlee et al. (2019) use sentiment analysis in order to examine the degree of negativity of messages that include gendered insults as well as adjectives used for reinforcing feminine stereotypes. The results show that by including insulting words that reinforce feminine stereotypes (especially references to physical characteristics) the degree of negativity of a message is significantly increased. Cryan et al. (2020) compare two methods for GS detection in job postings showing that a transformer (BERT) model outperforms a lexicon-based approach with adjectives and verbs that are potentially related to GS.

2.3 Sexist Hate Speech Detection

Waseem and Hovy (2016) provide the first corpus of tweets annotated with racism and sexism and use a logistic regression classifier with n-grams features for hate speech detection. There are also a few notable neural network techniques: LSTM (Jha and Mamidi, 2017) or CNN+GRU (Zhang and Luo, 2018). Chiril et al. (2020b) use a BERT model trained on word embeddings, linguistic features and generalization strategies to distinguish reports/denunciations of sexism from real sexist content that are directly addressed to a target.

Overall, as for stereotype detection, the work on automatic detection of sexist messages on social media is mainly supported by dedicated shared tasks that developed their own datasets, for example the AMI corpus mentioned above. These datasets (in English, Spanish and Italian) have also been used in the Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter shared task at SemEval 2019 (Basile et al., 2019). Best results were obtained with an SVM model using sentence embeddings as features (Indurthi et al., 2019). Lazzardi et al. (2021) conducted a study on this corpus to understand why participants obtained low scores on the identification of the particular type of misogynous behaviour against women (among which, stereotype, dominance, etc.) showing the difficulty of this task.

From the review of the literature, it is clear that GS is an under-explored area of research and approaches to automatic detection of stereotypes are very recent (either lexicon-based or deep learning models) and mainly deal with racist stereotypes. To our knowledge, no dedicated method for sexist hate speech classification taking into account GS has been developed. In this paper, we propose the first study that investigates how to improve sexist hate speech classification by using GS detection.

3 Data

3.1 Characterizing Gender Stereotypes

According to Haut Conseil à l'Égalité,[3] GS are schematic and globalizing representations that attribute supposedly "natural" and "normal" characteristics (psychological traits, behaviours, social roles or activities) to women and men. Deaux and Lewis (1984) define GS as having different and independent components (i.e., trait descriptors, physical characteristics, role behaviours and occupational status). These both definitions lead us to the definition of the following 3 categories of stereotypes. Note that when a stereotype is present, it can be expressed explicitly, implicitly (i.e., one can infer a content such as '*(all) women are...*') or it can be a denunciation/criticism of a GS.[4]

- **Physical characteristics** are related to physical strength or aspect. For example, the message *Short hair for a girl it's a bad idea* conveys the stereotype "Girls must have long hair".
- **Behavioural characteristics** are related to intelligence, emotions, sensibility or behaviour as in the denouncing tweet *Am I supposed to recognize myself in the "Just Fab" ad with a screaming hysterical bitch?*.
- **Activities** are activities, jobs, hobbies that are stereotypically assigned to women as in *Never marry a woman who cannot cook* which implies that a woman's place is in the kitchen, or *no woman understands football*.

Compared to existing datasets annotated for GS, ours offers a finer characterization (e.g., 2 categories in (Parikh et al., 2019) and only 1 in AMI), while capturing major stereotypes dimensions, as proposed in gender and communication science studies (Ellemers, 2018; Crawford et al., 2002).

[3] https://www.haut-conseil-egalite.gouv.fr/

[4] In order to better protect the privacy of the Twitter users, throughout this paper, instead of using direct quotations from the French tweets, we only provide their English translations.

3.2 Dataset for Gender Stereotype Detection

3.2.1 Stereo^{O}: The Original Dataset

As mentioned above, all existing datasets labelled with GS are dedicated to sexist hate speech detection and GS are considered as a form of sexism/misogyny. But a message containing a GS is not necessarily sexist and vice-versa (e.g., the message *"football is not for girls": it's over now!* contains the stereotype *girls cannot/must not play football* but the meaning conveyed by the whole message is not sexist. That is why we decided to rely on 2 different datasets for both sexism and GS detection tasks.

To build our dataset for GS detection, we used a non-annotated subset of 9,282 French tweets from the available corpus collected by (Chiril et al., 2020a) which contains 115,000 tweets collected using:[5] (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 (*#balancetonporc, #sexisme*, etc.). Given a tweet, its annotation consists in assigning it at least one of the following categories: physical characteristic, behavioural characteristic, activity and non-stereotype (the first 3 categories are not mutually exclusive). A tweet is annotated as "non-stereotype" when it does not contain a stereotype.

We hired two native French speaking annotators (one male and one female, both master's degree students in Linguistics, Communication and Gender) who after a training stage have annotated the corpus. 1,000 tweets have been annotated by both annotators so that the inter-annotator agreement could be computed (Kappa=0.79). Among the 9,282 annotated tweets, 91.47% contain no stereotype and 8.53% contain a stereotype. This results in a highly imbalanced dataset which size is relatively the same than in other datasets (e.g., 9% of the tweets contain a GS in the AMI corpora). Since only 10% of tweets get multiple labels, we decided to keep the predominant conveyed stereotype as the gold label for the experiments. Table 1 shows the distribution of the dataset, hereafter called Stereo^{O}.

[5] http://bit.ly/FrenchSexism

3.2.2 Stereo^{aug}: The Augmented Dataset

The corpus being quite small, especially the stereotype class, we decided to augment the training data to counter class imbalance. There are several strategies for data augmentation among which (see (Padurariu and Breaban, 2019) for an overview): oversampling (adding instances to the minority class with replacement (bootstrapping)), weighting the data during classification, adapting the loss function of the classification model, collecting more data or generating new instances similar to the ones belonging to the minority class. To generate new data, Ray et al. (2018) and Cho et al. (2019) use paraphrase generation in the domain of Spoken Language Understanding. Chawla et al. (2002) use the Synthetic Minority Oversampling Technique (SMOTE) which finds an instance similar to the one being oversampled and creates an instance that is a randomly weighted average of the original and the neighboring instance. Wei and Zou (2019) propose to extend data with simple operations: synonym replacement, random insertion, random swap, and random deletion. Hemker and Schuller (2018) use Natural Language Generation models for auto-generating new semantically similar instances based on the training data. However, the new instances with these methods may contain the same or similar words as the original instance but in a different order, which may result in generating instances that do not make sense to humans. In addition, these methods do not guarantee that the new generated instances belong to the same class as the original ones.

To avoid this, we propose a new approach for data augmentation based on sentence similarity. We use SentenceBERT, a modification of BERT that derives semantically sentence embeddings that can be compared using cosine-similarity (Reimers and Gurevych, 2019), to extend our training dataset with the most similar sentences from two sources: (S1) **New tweets in French** collected with a small set of keywords usually used in stereotypes about women: *moche (ugly), fesses (butt), jupe (skirt), bavarde (gossipy), dépensière (spendthrift), dévouée (devoted), infirmière (nurse), poupée (doll)*. These keywords are different from those used for the initial data collection; and (S2) **New tweets from existing multilingual datasets** annotated for stereotypes. Since there is no other available resource in French, we tried to extend our initial training corpus in two ways:

Non Stereotype 8490	Stereo^{O} 792			Stereo^{aug} Initial French: 792 Eng IberEval: 1,914 / New Fr: 2,241		
	physical	behaviour	activity	physical	behaviour	activity
	170	210	412	689	473	1224

Table 1: Stereotype corpus distribution in the initial and augmented datasets.

(a) Augmenting with multilingual instances annotated as stereotypes from AMI (English, Italian, Spanish) and the English sexism corpus (Parikh et al., 2019). This strategy did not lead to good results in the following experiments;

(b) Augmenting with the most similar instances to the ones labelled as stereotype in our corpus as given by SentenceBERT. To this end, we consider the aforementioned corpora, as well as (Waseem and Hovy, 2016). The dataset augmented via similarity from the English IberEval lead to best results. This is the one we use hereafter (**Stereo**^{aug}).

For both sources of augmentation (i.e., (S1) and (S2)), a threshold T was set experimentally and the most similar instances from IberEval dataset and new collected tweets were automatically labelled as *stereotype* and added to our training dataset.[6] [7] This allows to select similar instances in terms of vocabulary (cf. (1)) but also of syntactic patterns (cf. (2)).

(1) Initial tweet: *I admit that the kitchen is the uncontested territory of women.*
Similar English tweet (T=0.459): *#YesAllWomen belong in the kitchen*

(2) Initial tweet: *Why is there always a window in the kitchen? So that women can have a point of view.*
Similar English tweet (T=0.496): *Why do women get married in white? So they match the kitchen appliances.*

Finally, Stereo^{aug} is now composed of 4,891 tweets which represents an augmentation of about 45% of the initial corpus (see distribution in Table 1). For the experiments, all new augmented instances are added to the train while the initial dataset have been divided into train (80%) and test (20%) sets. The test set being the same in all configurations and composed only of initial tweets from Stereo^{O}.

4 Gender Stereotype Detection

4.1 Models

Our objectives are twofold: (1) Investigate the effectiveness of sentence similarity as a data augmentation strategy; (2) Identify the most appropriate deep learning architecture able to capture the linguistic characteristics of GS in short messages. To this end, we propose several models relying on different contextualized pre-trained models as input: either FlauBERT[8] (Le et al., 2020) or Multilingual BERT[9] (Devlin et al., 2019). The FlauBERT based models were trained on the original dataset (i.e., Stereo^{O}), while the multilingual BERT based models were trained on the augmented dataset (i.e., Stereo^{aug}). In this way, we are comparing different methods employed for stereotype detection on both the original and augmented datasets.

FlauBERT$_{\text{base}}$/BERT$_{\text{base}}$. These are our baselines that respectively use FlauBERT-Base Cased and BERT-Base Multilingual Cased without any additional inputs. Both models were implemented using the HuggingFace library (Wolf et al., 2019).

FlauBERT$^{L}_{\text{base}}$. This model is similar to FlauBERT$_{\text{base}}$, but it uses focal loss (Lin et al., 2017) instead.[10] Our aim here is to compare with one of the most effective approach for handling imbalanced data (Cui et al., 2019).

FlauBERT$_{\text{lex}}$/BERT$_{\text{lex}}$. In order to force the classifier to learn from generalized concepts rather than words which may be rare in the corpus, we adopt several replacement combinations extending Badjatiya et al. (2017)'s and Chiril et al. (2020b)'s approach. We used a publicly avail-

[6] $T = 0.45$ for the IberEval dataset and $T = 0.5$ for the newly collected French data as the number of similar instances returned was higher.

[7] When performing the augmentation strategy for instances with multiple labels, if the same instance was retrieved for more than one category, it was not included in the augmented dataset (this is the reason why in Table 1 the number of instances in Stereo^{aug} for the binary classification is different than for multi-label classification).

[8] Note that when choosing the best BERT variant for Stereo^{O} we experimented with different models: multilingual BERT, CamemBERT (Martin et al., 2019) and FlauBERT. FlauBERT outperformed the other two models.

[9] As Stereo^{O} is multilingual (i.e., it contains instances in both French and English) we had to use BERT multilingual.

[10] Results with dice loss (Li et al., 2020) were lower.

able French lexicon comprising 130 gender stereotyped words[11] that we grouped according to our 3 categories (*physical characteristics*, *behavioural characteristics*, *activities*) and replaced these words/expressions when present in tweets by their category. Note that only 1% of these words overlap with the ones used to collect the initial and extended datasets. When applied on English inputs, we automatically translated the words by aligning French and English FastText word vectors (Conneau et al., 2017) and selecting the nearest neighbor in the target space.

FlauBERT$_{\text{ConceptNet}}$/BERT$_{\text{ConceptNet}}$. Instead of relying solely on manually built lists of words, we try to automatically extend them with words extracted through ConceptNet (Speer et al., 2017), a multilingual knowledge graph for natural language words or phrases in their undisambiguated forms. Although similar knowledge bases exist (e.g., BabelNet (Navigli and Ponzetto, 2012)), our choice is motivated by the fact that for a given word, ConceptNet is focusing on common-sense relationships to other words, as opposed to BabelNet, which focuses on dictionary definitions of words (i.e., WordNet-style synsets). In addition, ConceptNet has a larger coverage for French. Lexicon extension works as follows:[12] Given a word in the French lexicon, we extend it via the relations SimilarTo and Synonym.[13] For example, for *bavarde (talkative)*, the retrieved words includes *jacasse (chatter)* and *commère (gossip girl)*. After following this strategy, we obtained a total of 725 entries in French (used for FlauBERT) and 1,993 entries in French and English (used for BERT).

FlauBERT$_{\text{label_emb}}$/BERT$_{\text{label_emb}}$. Our stereotype categories being relatively informative, another way to force the classifier to infer the correct link between a given message and the GS it may evoke is to leverage additional information as given by the labels themselves. We therefore propose to use label embedding (Wang et al., 2018), a technique that embeds both class labels and the text into a joint latent space, where the model can be trained to cross-attend the inputs and labels in order to improve the model performance. Our models are similar to (Si et al., 2020) who consider the joint representation of the tweet and its corresponding class token and incorporate label embeddings into the self-attention modules. The label embeddings for the class stereotype are initialized as the average of the corresponding keyword embeddings (here, we consider the words in the lexicon as keywords representative for the class stereotype), while the label embedding for the non-stereotype class is initialized at random. For $\texttt{Stereo}^{aug}$, the English keywords were obtained in the same manner as for BERT$_{\text{lex}}$.

4.2 Results and Discussion

All the proposed models have been evaluated on $\texttt{Stereo}^{O}$ test set while the hyperparameters were tuned on the validation sets (20% of the training dataset), such that the best validation error was produced. Stereotype detection, and GS in particular, being a new task, there is no strong state of the art models to compare with apart Sánchez-Junquera et al. (2021) and the winner system at HaSpeeDe2 by Lavergne et al. (2020) for binary stereotypes detection against immigrants and the one by Cryan et al. (2020) for binary gender bias classification in job postings. Both models are based on pre-trained contextualized embeddings which have been fine tuned on the task without accounting for any prior linguistic knowledge about GS. These models are thus similar to our **FlauBERT$_{\text{base}}$** and **BERT$_{\text{base}}$**. Since current studies consider GS as a type of sexism/misogyny, we also compare with the best performing models for sexist hate speech detection: **CNN$_{\text{FastText}}$** (Karlekar and Bansal, 2018) that uses FastText pre-trained French word vectors (with the dimension of 300), **CNN-LSTM** (Karlekar and Bansal, 2018; Parikh et al., 2019) based on the previous CNN model by adding an LSTM layer[14] except that we used word-level embeddings instead of character/sentence-level as the results were lower, and finally, **BiLSTM with attention** (Parikh et al., 2019).

Table 2 presents the results for the binary GS detection task in terms of macro-averaged F-score (F), precision (P) and recall (R) with the best results presented in bold. We observe that best baselines are without surprise FlauBERT$_{\text{base}}$ and BERT$_{\text{base}}$ and more importantly, that data augmentation via sentence similarity as given by SentenceBERT is very effective. Indeed, the model trained on $\texttt{Stereo}^{aug}$ achieves better results

[11] http://bit.ly/FrenchSexism

[12] We also tried extending these lexicons by selecting only three seed words from each of the lexicon's categories, however we noticed that the results tend to decrease.

[13] Extension via RelatedTo was not conclusive.

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

than the one trained on $\texttt{Stereo}^{O}$, outperforming FlauBERT$^{L}_{\text{base}}$, the model designed to handle class imbalance in the original dataset. Another important finding is that all the models that incorporate GS knowledge improve over the baselines, the best strategy being the one based on ConceptNet. Also, the results for label embeddings are close to the one based on manual lexicon of GS. These results suggest that in the absence of a lexicon, label embeddings could be a valid strategy.

Overall, we can conclude that coupling GS information as encoded in external lexicons (either manually built or extended) with contextualized representation of words is a good strategy, enabling the classifier to learn from generalized concepts rather than words themselves. However, even if this strategy relies on a manual list of seed words in a given language, we show that it is generic enough since it is both (a) *language independent* thanks to knowledge graphs such as ConceptNet that was able to capture word similarity in a multilingual context, and (b) *target independent and transferable to other languages* because lists of representative stereotype words targeting other social groups can be easily built by automatically extending existing compiled lists proposed in the literature (e.g., (Garg et al., 2018) for ethnic stereotypes and HurtLex (Bassignana et al., 2018) for negative stereotypes).

CLASSIFIER	P	R	F
CNN‡	0.619	0.630	0.624
CNN+LSTM‡	0.572	0.622	0.595
BiLSTM$_{\text{attention}}$‡	0.589	0.593	0.590
FlauBERT$_{\text{base}}$‡	0.656	0.659	0.658
FlauBERT$^{L}_{\text{base}}$	0.672	0.667	0.669
BERT$_{\text{base}}$‡	**0.734**	0.706	0.719
FlauBERT$_{\text{lex}}$	0.674	0.693	0.683
BERT$_{\text{lex}}$	**0.734**	0.718	0.725
FlauBERT$_{\text{ConceptNet}}$	0.711	0.704	0.708
BERT$_{\text{ConceptNet}}$	0.726	**0.731**	**0.729**
FlauBERT$_{\text{label_embeddings}}$	0.685	0.680	0.682
BERT$_{\text{label_embeddings}}$	0.729	0.717	0.724

Table 2: Results for the most productive strategies for binary classification. ‡: baseline models.

The macro F-scores per class as given by our best model BERT$_{\text{ConceptNet}}$ are 0.725 for Activity, 0.693 for Physical and 0.583 for Behaviour, while the macro score for 4 classes classification including the non stereotype is 0.510. A manual error analysis shows that misclassification cases are due to 2 main factors: the presence of a GS along with its contrary (denouncing tweets) leading to false negatives (58% of missclassifications) as in (3), and the presence of many words designating or describing women along with words usually used in GS leading to false positives as in (4).

(3) *Justin Trudeau is shirtless: he breaks the rules.* ***A woman wears a short dress: it's unbearable****. In France,* ***women have the right to dress as they want****.*

(4) *I don't understand people who support several clubs. You love only one* ***woman****, you have only one* ***mother****. It's the same for* ***football****, you love only one club.*

5 GS for Sexist Hate Speech Detection

5.1 Models

We aim to show how GS prediction (considered as an auxiliary task) can be used for sexism detection (the main task). To this end, we used the only available resource in French from (Chiril et al., 2020a): 11,834 tweets annotated with the *sexist* tag if the tweet conveys a sexist content and *non-sexist* if not, the distribution being 34.2% for the positive class and 65.80% for the negative one. 20% of the data has been used for testing our models. It is important to note that as there is no overlap between this dataset and the GS one, this will prevent the models for sexism detection (which will integrate stereotype prediction) to be biased.

Several strategies for injecting the stereotype information in the sexism detection task were explored, ranging from using the predictions of the best stereotype model to multitask approaches (Ruder, 2017). To this end we compare with: (1) the only existing model for French for detecting sexist hate speech (Chiril et al., 2020b), and (2) existing models that consider stereotypes as an auxiliary task to improve hate speech classification. Lavergne et al. (2020) is the only team in the recently shared task HaSpeeDe 2 that considers the interaction between hate speech towards immigrants and racial stereotype detection by employing a multitask learning approach.

BERT$_{\text{gen}}$. It takes the best model proposed in (Chiril et al., 2020b) which is based on BERT and trained on word embeddings, linguistic features (surface and opinion features) and generalization strategies (replacement of places and persons by an hypernym).

BERT$_{\text{tag}}$. It uses the predictions of the best performing model for stereotype detection (i.e., BERT$_{\text{ConceptNet}}$ trained on the augmented dataset)

for adding at the end of each tweet a tag indicating the presence of stereotypes ($\text{BERT}_{\text{tag\_binary}}$) or the type of stereotype ($\text{BERT}_{\text{tag\_type}}$).

$\text{MT}_{\text{Lavergne}}$ (Lavergne et al., 2020). It is based on a BERT multitask architecture trained on a dataset annotated for both the presence of hate speech and stereotypes. However, in our case, since we rely on two different datasets (one for each task), we used the stereotype predictions of the best performing stereotype model (i.e., $\text{BERT}_{\text{ConceptNet}}$) to automatically label the sexism dataset with stereotype information.

AngryBERT (Awal et al., 2021). This model was specifically designed to address the problem of imbalanced datasets by jointly learning hate speech detection with emotion classification and target identification as secondary tasks. It has been shown to outperform many strong existing multitask models, including MT-DNN (Liu et al., 2019). In our case, the primary task of AngryBERT is sexism detection while the second being the detection of stereotypes. In addition to this initial configuration (**$\text{AngryBERT}_{\text{base}}$**), four models are newly proposed, depending on both (i) the number of labels to predict in the auxiliary task, and (ii) the dataset on which the generalization with hypernyms is performed. Chiril et al. (2020b) showed that on their sexism dataset the generalization strategy performs well. In addition, we observed that a similar generalization can be employed for our task with good results. Based on these observations we are analyzing whether this generalization approach should be adopted in the sexism (i.e., **$\text{AngryBERT}_{\text{sexism}}$**) or in the stereotype dataset (i.e., **$\text{AngryBERT}_{\text{stereo}}$**).[15] In addition, as the GS dataset does not contain only instances annotated as *stereotype* vs. *non-stereotype*, but also different categories, we are analyzing whether the auxiliary task should be binary (i.e., **AngryBERT^2**) or multiclass (i.e., **AngryBERT^4**). For all the settings, the auxiliary task was trained on the augmented multilingual dataset and the generalization relies on ConcepNet, as it performed the best (cf. Section 4.2).

5.2 Results and Discussion

Table 3 presents the multitask and the baselines results. We observe that injecting stereotypes labels as given by the automatic classifier (i.e., BERT_{tag}) outperforms both $\text{MT}_{\text{Lavergne}}$ and $\text{AngryBERT}_{\text{base}}$ the two multitask baselines. In particular, predicting the types of stereotypes is the most productive when compared to presence identification (F-score 0.796 vs. 0.776). However, when GS information is predicted jointly with sexist labels, the results tend to decrease for all AngryBERT configurations except for $\text{AngryBERT}^2{}_{\text{sexism}}$ and $\text{AngryBERT}^4{}_{\text{sexism}}$ in which we performed ConcepNet generalization on the sexism dataset only. Here again, GS types are the best with an F-score of 0.827, significally beating our strong baseline BERT_{gen} ($p < 0.05$ using the McNemar's Test statistic).

A closer look into the results per class shows that $\text{AngryBERT}^4{}_{\text{sexism}}$ was able to better predict sexist content (F-score=0.805 vs. 0.773 for BERT_{gen}). This suggests that GS information is definitively helpful for sexist content detection when it is injected as additional knowledge on top of the primary task.

An error analysis shows that 59% of missclassified instances are false negatives (sexist tweets detected as non sexist) and among them only 7% contain a GS (with a manual observation). This suggests that the majority of these sexist instances cannot benefit from the GS auxiliary task, confirming that sexist content does not necessarily entail the presence of stereotypes, as in (5).

(5) *Ségolène Royal is lucky, they don't eat turkey for #ThanksGiving at the Poles! #TheSurvivorTurkey*.

Among the false positives (non sexist tweets detected as sexist), 93% are predicted as non stereotype and a manual observation confirms that only 4% contain a GS. This means that the classification errors are due to the sexism classifier. When looking at these instances, we note that 57% contain hashtags usually dedicated to sexism which are misused as in (6).[16]

(6) *Why isn't there any pastry chef who puts strange food like tomato, guacamole #TopChef* ***#BalanceTonPorc***

As shown with the above examples, error classifications are often due to humor, jokes, irony or puns, meaning that accounting for these phenomena for hate speech detection is still an open problem.

[15] Note that we do not perform the generalization in both datasets as to not introduce bias.

[16] Note that the distribution of keywords/hashtags is very similar in both non-sexist/non-stereotype and sexist/stereotype tweets which means that the presence of hashtags have little impact on the classification performances .

CLASSIFIER	P	R	F
BERT_{gen} ‡	**0.865**	0.787	0.824
$\text{BERT}_{\text{tag\_binary}}$ ‡	0.821	0.736	0.776
$\text{BERT}_{\text{tag\_type}}$ ‡	0.835	0.761	0.796
$\text{MT}_{\text{Lavergne}}$ ‡	0.803	0.749	0.775
$\text{AngryBERT}_{\text{base}}$ ‡	0.725	0.727	0.726
$\text{AngryBERT}^{2}{}_{\text{stereo}}$	0.730	0.728	0.729
$\text{AngryBERT}^{4}{}_{\text{stereo}}$	0.733	0.737	0.735
$\text{AngryBERT}^{2}{}_{\text{sexism}}$	0.836	0.813	0.824
$\text{AngryBERT}^{4}{}_{\text{sexism}}$	0.839	**0.816**	**0.827**

Table 3: Results for sexist classification. ‡: baselines.

6 Conclusion

In this paper, we proposed the first approach for gender stereotype detection in tweets as well as several deep learning strategies to inject appropriate knowledge about how stereotypes are expressed in language into sexism hate speech classification. Our main results are: (1) a new dataset for GS detection, (2) a method to counter class imbalance based on sentence similarity from multilingual external datasets, (3) different strategies to incorporate GS triggers as input into the learning process based on automatically extended lexicon via a multilingual knowledge graph, and finally, (4) an empirical evaluation of the positive impact of multiclass GS detection on improving hate speech against women based on multitask architectures, beating several strong state of the art baselines. Although our approach is specific to gender stereotyping, we believe it is generic enough to detect other types of stereotypes like the ones related to racism through the use of other resources (e.g., ConceptNet, BabelNet, Hurtlex, etc.), without presuming performances.

GS is an understudied problem and we believe it should not only be viewed as a type of sexism/misogyny but considered instead as an independent task to be used in other applications as well. Among them, education is a promising future direction for selecting which digital media/books are being given to children, as previous research has indicated that the stereotypes children encounter in their environment can impact their motivational dispositions and attitudes. In the future, we plan on addressing these issues, as well as developing approaches for leveraging the GS information in other datasets annotated for sexism.

Ethical Approval. This article does not contain any studies with human participants carried out by any of the authors. In addition, the data that was used is composed of textual content from the public domain taken from datasets publicly available to the research community. These datasets also conform to the Twitter Developer Agreement and Policy that allows unlimited distribution of the numeric identification number of each tweet. For the GS corpus, the data have been annotated with respect to certain types of stereotypical language, however, we are not making any claims about the authors of the tweets, neither share a large numbers of tweets from the same users. Additionally, if any of the users want to opt out from having their data being used for research, they can request that they be removed from the dataset by sending an email to the authors of this paper. This work offers several positive societal benefits. Sexism is a well-known problem, and countering it via automatic methods can have a big impact on people's lives. This challenge is meant to spur innovation and encourage new developments for both sexism detection and stereotype detection which can have positive effects for an extremely wide variety of tasks and applications. With these advantages also come potential downsides.

The GS dataset is not intended to be used for collecting user information which could potentially raise ethical issues. Relying on models flagging posts as sexist/conveying stereotypes based on user statistics might be biased towards certain users which eventually could limit freedom of speech on the platform.

Acknowledgements

We would like to thank the annotators: Mathilde Espercé and Frédéric Saudemont. We also thank the anonymous reviewers as well as the meta reviewers for their useful comments that helped improve this paper. This work has been carried out in the framework of the STERHEOTYPES project funded by the Compagnia San Paolo 'Challenge for Europe', as well as the INTACT project funded by the AAP CNRS - INHESJ 2020.

References

Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. The nature of prejudice.

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 Natu-*

ral Language to Information Systems, NLDB 2018, pages 57–64.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. 2021. Angrybert: Joint learning target and emotion for hate speech detection. *arXiv preprint arXiv:2103.11800*.

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.

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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Denise R Beike and Steven J Sherman. 2014. Social inference: Inductions, deductions, and analogies. *Handbook of social cognition,*, pages 209–285.

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.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. 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), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020a. An annotated corpus for sexism detection in french tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1397–1403.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020b. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066.

Eunah Cho, He Xie, and William M. Campbell. 2019. Paraphrase generation for semi-supervised learning in NLU. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 45–54, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Matthew T Crawford, Steven J Sherman, and David L Hamilton. 2002. Perceived entitativity, stereotype formation, and the interchangeability of group members. *Journal of personality and social psychology*, 83(5):1076.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. 2020. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kay Deaux and Laurie L Lewis. 1984. Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of personality and Social Psychology*, 46(5):991.

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.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring Human Gender Stereotypes with Word Association Test. In *Proceedings of the EMNLP-IJCNLP*.

Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology*, 69:275–298.

Diane Felmlee, Paulina Inara Rodis, and Amy Zhang. 2019. Sexist slurs: Reinforcing feminine stereotypes online. *Sex Roles*, pages 1–13.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of IberEval@SEPLN*.

Susan T Fiske. 1998. Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, 2(4):357–411.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Chiara Francesconi, Cristina Bosco, Fabio Poletto, and Manuela Sanguinetti. 2019. Error analysis in a hate speech detection task: The case of haspeede-tw at evalita 2018. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Ruben García-Sánchez, Carmen Almendros, Begona Aramayona, Soria-Oliver Maria Martín, Maria Jesus, Jorge S. López, and José Manuel Martínez. 2019. Are Sexist Attitudes and Gender Stereotypes Linked? A Critical Feminist Approach With a Spanish Sample. Frontiers in psychology. *Front Psychol. 2019;10:2410*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Elizabeth L Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing… or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

Konstantin Hemker and Bjorn Schuller. 2018. Data augmentation and deep learning for hate speech detection. *Imperial College London*.

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.

Eric Lavergne, Rajkumar Saini, György Kovács, and Killian Murphy. 2020. Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection. In *7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020*, volume 2765. CEUR-WS.

Silvia Lazzardi, Viviana Patti, and Paolo Rosso. 2021. Categorizing Misogynistic Behaviours in Italian, English and Spanish Tweets. *Procesamiento del Lenguaje Natural (SEPLN), num. 66*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Walter Lippmann. 1946. *Public opinion*, volume 1. Transaction Publishers.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies.

In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 92–105.

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.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Cristian Padurariu and Mihaela Elena Breaban. 2019. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 23rd International Conference KES2019.

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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

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.

Avik Ray, Yilin Shen, and Hongxia Jin. 2018. Robust spoken language understanding via paraphrasing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@ EVALITA2020: Overview of the Evalita 2020 hate speech detection task. In *Proceedings of EVALITA*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of LREC*.

Shijing Si, Rui Wang, Jedrek Wosik, Hao Zhang, David Dov, Guoyin Wang, and Lawrence Carin. 2020. Students need more attention: Bert-based attention model for small data with application to automatic patient message triage. In *Machine Learning for Healthcare Conference*, pages 436–456. PMLR.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Javier Sánchez-Junquera, Berta Chulvi, Paolo Rosso, and Simone Ponzetto. 2021. How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. *Applied Sciences, 11(8), 3610*.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia. Association for Computational Linguistics.

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.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

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, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.