Misogyny in Twitter
Can a woman's career have an impact on the attacks she receives online?

Alessio Cascione[•], Aldo Cerulli[•], Alessandra Cicciari[◇], Umberto Lamia[◇]
[•] Department of Philology, Literature and Linguistics, University of Pisa
[◇] Department of Computer Science, University of Pisa
{a.cascione, a.cerulli1, a.cicciari, u.lamia}
@studenti.unipi.it

Abstract

This report aims at expanding the studies in literature about automatic misogyny identification in social media by investigating whether and in which measure women occupational fields play a role in the attacks they receive online. The described study is threefold: it expands the chosen starting data-set comprising misogynistic and non-misogynistic tweets by adding newly extracted tweets provided with the indication of victims' occupation; assesses a state-of-the-art classifier's performance on the misogyny identification task, exploiting it to perform automatic annotation of the mentioned newly extracted tweets with respect to different categories of misogynistic behaviors; retrieves insights regarding the relationship between women's jobs and the kinds of misogyny directed against them.

1 Introduction

Misogyny on social media is a noted contemporary phenomenon. In contrast to general hate speech, social media contents that can be defined as misogynistic are specifically addressed to women.

It is dutiful to do a clarification about a distinction that came to light during the analysis carried out for the purposes of our investigation: the difference between a 'misogynistic language' and a 'misogynous behaviour actually aimed at offending a woman'. The former concerns the use of some words in a way that the text results to be misogynous but not necessarily addressed to a woman (e.g. a man, a political party, a company); the latter is configured as the use of misogynous expressions with the precise objective of attacking a woman. In our study, both cases are taken into consideration in different phases and equally labeled as misogynous content.

The detection of misogyny on social media is crucial for preventing and reducing the production and diffusion of disrespectful contents and possibly punishing their authors. To pursue this goal, more and more sophisticated and performing NLP-techniques are being developed by researchers all over the world. Among the other contributions, our inspiring example of the use of NLP in this direction is given by the 'AMI' (*Automatic Misogyny Identification*) shared task. It was presented for the first time in 2018 in occasion of the IberEval[1] and EVALITA[2] evaluation campaigns and then proposed again in the following edition of EVALITA in 2020[3]. Our study is aimed at deepening the 'AMI' task by examining how the presence of misogyny varies across different job categories, in order to understand in which fields women are more likely to be subject to it.

Contributions. To the best of our knowledge, this is the first study that (i) introduces an English data-set of misogynous tweets annotated with the *job category* to which the addressed women belong, resulting by the expansion of the training set provided by the promoters of EVALITA 2018 'AMI' task by means of the addition of several (especially) misogynous tweets; and (ii) reports classification task results performed with state-of-the-art algorithms.

In what follows, we first present the data-set on which our work is based as well as the operations of exploration and expansion we carried out in order to collect more misogynous observations with respect to different kinds of jobs (Section **??**). In Section 3 we describe the training process of different types of binary and multi-class classification algorithms and how the most performing one was used to automatically annotate the new misogynous tweets with respect to their typology of misogynous behavior; Section 4 analyses the final data-set ob-

[1] https://amiibereval2018.wordpress.com/ (06/02/23)
[2] https://amievalita2018.wordpress.com/ (06/02/23)
[3] https://amievalita2020.github.io/ (06/02/23)

tained by combining the new misogynous tweets with those belonging to the original data-set for which it was possible to identify the occupational area of the offended woman. In the conclusions, we sum up the most significant findings of the project and propose further improvements in several directions.

2 Data-set exploration and enlargement

As already mentioned, the starting point of our study was the analysis of the data-set provided for the development of the EVALITA 2018 'AMI' shared task (Fersini et al., 2018), which in turn is an improved version of the one created for the same task in occasion of the IberEval 2018 campaign. The data-set is composed by a training set (4,000 tweets) and a test set (1,000 tweets). Both sets are structured into five columns: the *id* uniquely identifies each tweet; *text* reports the content of the tweets; *misogynous* has value '1' if the tweet is misogynous, 0 otherwise; *misogyny category* can assume one among five strings defining as much misogynistic behaviors (i.e., *Derailing*, *Discredit*, *Dominance*, *Sexual harassment*, *Stereotype*); and *target* (*Active* or *Passive*). Figure 1 shows the heading of the data-set and some examples of tweets annotated according to the mentioned levels.

id	text	misogynous	misogyny_category	target
1	Please tell me why the bitch next to me in the...	1	dominance	active
2	@emmasharp003 @Ldrake48Lee Bitch shut the fuck up	1	dominance	active
3	@abzdafab Dear cunt, please shut the fuck up.	1	dominance	active

Figure 1: Structure of the original data-set

Analyzing the distribution of tweets with respect to the 'misogynous' variable (Figure 2), it resulted that the number of non-misogynous tweets is quite higher (2215 against 1785).

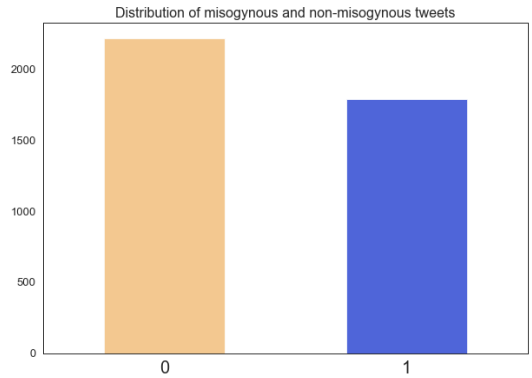

Figure 2: Misogynous and non-misogynous tweets

From the beginning, we decided to not use the feature *target*, since not useful for our purpose; on the other hand, we proceeded with the analysis of the different kinds of misogynistic behaviors. As Figure 3 shows, the distribution of the column *misogyny category* in the original data-set is very imbalanced. Moreover, since the differences between some categories are fine-grained, it is very difficult to catch them. We tried to find a remedy for it using the BERTweet model (more details will be given in Section 3.2).

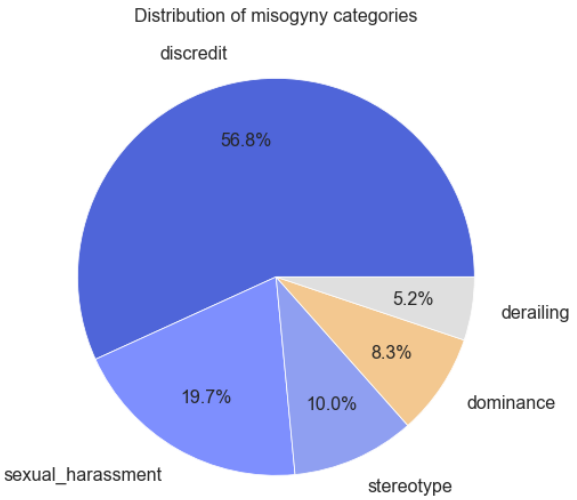

Figure 3: Distribution of tweets with respect to 'misogyny_category'

2.1 Data-set annotation with professions

As previously claimed, the main purpose of our project is to study the misogyny phenomenon in Twitter with respect to the profession of the victims. Since an indication about women jobs was not included in the original data-set, we decided to add a new level of annotation to it - i.e., a column 'occupation' - specifying for each tweet the kind of job held by the victim. In order to do that, we firstly identified five different classes - *Arts, Show, Politics and Activism, Science and Culture, Sport* - that, in our opinion, could cover most possible jobs. This is a first attempt to perform such a categorization, even if we are well aware that the coverage of our selected fields is not exhaustive. For this reason, during the annotation phase, we had to introduce two other labels to classify tweets referred to women outside these fields: 'other', to encompass a heterogeneous group of cases - i.e., video games, movies and TV series characters, saleswomen, employees and even the Virgin Mary - and 'no_woman' to annotate all the tweets

exhibiting a 'misogynistic language' against men, companies - e.g., the NASA - and political parties - e.g., the PMLN (*Pakistan Muslim League*). The latter captures the distinction, already explained in the Introduction, between 'misogynistic language' and a 'misogynous behavior'. Moreover, we used the label 'generic_g' to classify tweets addressed to women in general or wide groups of them - e.g., 'women in Austin', 'Women in their 30's' - and 'generic_r' for tweets that referred to specific women, whose identity and/or job was impossible to be detected.

As shown in Table 1, the overwhelming majority of tweets are generic (of both types), while only 218 tweets were assigned one of the five chosen fields (in bold). They are obviously insufficient to carry out meaningful analyses. Therefore, we decided to expand the data-set.

Labels	Number
generic_r	760
generic_g	696
pol_act	97
show	87
no_woman	35
sci_cul	21
other	18
sport	7
arts	6

Table 1: Number of tweets per label

The process of annotation allowed us to identify several inconsistencies on the original data-set. In addition to finding many misogynous tweets addressed not to women - those we labelled with 'no_woman' -, we identified a lot of non-misogynistic tweets classified as misogynous. We decided to (1) keep the former into consideration only for the training of classification algorithms, discarding them for the final analyses and (2) change the annotation of the latter according to the way non-misogynous contents are annotated in the original data-set. Furthermore, many tweets reported song titles and/or lyrics. We opted to delete them since we could not establish if their goal was actually to offend some woman.

2.2 Data-set expansion

Reading and annotating the tweets we found out that there are some words that occur with very large frequencies. This evidence allowed us to assume that the creators of the data-set had probably used them to download tweets to be included in the data-set. Therefore, for each misogyny category, we extracted the most frequent words. This operation was firstly aimed at having a general understanding of the content conveyed by the tweets, but above all at collecting the keywords used by the authors and reuse them to download new tweets in order to give consistence to the extracted data with respect to the original corpus. The obtained words have been then used to build an unique list containing the most common expressions over the entire data-set[4].

Secondly, we selected a number of popular women representative of the five occupational fields mentioned before (see Table 2).

Occupation	Woman
Arts	Marina Abramović Neri Oxman Tracy Reese Ariana Richards Stella McCartney Annie Leibovitz Kathryn Bigelow Cindy Sherman Barbara Kruger Jenny Saville Shirin Neshat Jil Sander Julie Mehretu Yayoi Kusama Aurora James Wangechi Mutu
Politics and activism	Theresa May Theresa May Michelle Obama Alexandria Ocasio-Cortez Hillary Clinton Ursula Von der Leyen Greta Thunberg Nancy Pelosi Malala Yousafzai
Science and culture	J.K. Rowling Tara Westover Samantha Cristoforetti Ilaria Capua Katie Mack Raychelle Burks
Show	Kendall Jenner Ariana Grande Selena Gomez Scarlett Johansson Ellen Degeneres Kim Kardashian
Sport	Serena Williams Allyson Felix Hope Solo Carmelita Jeter Alex Morgan Natalie Coughlin Misty May-Treanor

Table 2: Selected women per occupational field

Afterwards, we wrote a scraping function that,

[4]The list comprises the following words: bitch, women, shut, fuck, men, womensuck, fucking, skank, cunt, hoe, rape, whore, dick, ass, sexual, assault, victim, yesallmen, pussy, cock, suck, shit, stupid, bitches, slut, hysterical, fat, ugly.

taking a positive integer (*N*) as input, downloads *N* tweets that contain an input string reporting the name of a woman in the form of a hashtag, a mention and/or directly the sequence 'Name and Surname' (e.g., #samanthacristoforetti, @AstroSamantha and/or 'Samantha Cristoforetti'). We extracted both misogynous and non-misogynous tweets, maintaining a higher number of the former given our study is focused on them.

Overall, we collected 990 new tweets, divided into 760 misogynous e 230 non-misogynous. Table 3 shows the distribution of new tweets with respect to kind of job.

Occupation	Misogynous	Non-misogynous	Total
Arts	145	76	221
Politics and activism	166	76	242
Science and culture	140	46	186
Show	73	6	79
Sport	236	26	262
Total	760	230	990

Table 3: Misogynous and non-misogynous extracted tweets per occupational field

3 Classification

The following section is dedicated to the assessment of different classification approaches to the misogyny identification problem. We first show the performance of "classic" approaches for general classification problems and then focus our attention on state-of-the-art models in NLP tasks (BERT and derived models). For both standard and transformer approaches, a corrected version of EVALITA 2018 data-set has been pre-processed with TweetNormalizer, proposed alongside BERTweet in (Nguyen et al., 2020) [5], which applies TweetTokenizer from the NLTK toolkit[6], map emojis into text strings and substitute user mentions and web/url links with @USER and HTTPURL.

3.1 Standard models

For both classification tasks, we take into account three different model classes implemented in Scikit-learn for Python (Pedregosa et al., 2011): we will focus on performances given by Logistic Regression (LR), Random Forest (RF) and Multi Layer Perceptron (MLP) solvers trained on the corrected data-set. Concerning the general methodology adopted for model selection and assessment for each class, we specify that for each kind of solver we performed a gridsearch using a stratified *k*-fold cross validation approach with *k*=5 in order to identify the best parameters for that particular class of models. As scoring criteria to consider for the extraction of the best model from each respective gridsearch, accuracy has been chosen for the binary case and macro-averaged F1 score for the multi-class one. Models reported in Table 4 and 5 are the ones with the highest mean score over the validation folds according to their respective gridsearch. We present evaluations first taking in isolation linguistic features, sentence embeddings[7] and uni/bi/tri-grams of lemmas and single named entities, then results obtained merging the features together in order to expand the feature space for each tweet[8].

Feature Space	Logistic Regression		Random Forest		Multi-Layer Perceptron	
	Mean VL	TS	Mean VL	TS	Mean VL	TS
N-grams	77.37% ± 1.23%	63.80%	78.36% ± 1.61%	64.00%	76.86% ± 0.83%	61.50%
Embeddings	78.72% ± 1.39%	69.00%	76.70% ± 0.93%	67.90%	78.75% ± 1.68%	67.60%
Linguistic	58.15% ± 2.16%	53.70%	59.88% ± 0.98%	56.70%	60.78% ± 0.51%	57.60%
Complete	80.47% ± 1.27%	69.60%	78.03% ± 0.87%	66.80%	77.53% ± 2.29%	69.90%

Table 4: Binary-classification task

Feature Space	Logistic Regression		Random Forest		Multi-Layer Perceptron	
	Mean VL	TS	Mean VL	TS	Mean VL	TS
N-grams	37.49% ± 1.91%	46.13%	36.15% ± 2.61%	45.82%	42.59% ± 3.02%	45.61%
Embeddings	42.27% ± 3.34%	36.78%	25.56% ± 1.89%	19.35%	48.57% ± 2.98%	45.04%
Linguistic	14.50% ± 0%	12.74%	20.88% ± 1.90%	16.06%	16.44% ± 0.92%	11.23%
Complete	45.60% ± 4.47%	45.84%	24.85% ± 2.63%	20.34%	44.97% ± 3.69%	50.33%

Table 5: Multi-classification task

All the classifiers seem to show very similar performances in the binary context, while stronger differences arise in the multi-class scenario. We will report here the parameters used in the model selection for each solver class and focus on the ones chosen for the best overall result, considering the best test set performance across every feature space considered. With LR, we tried for three different solvers(lbfgs,newton-cg,liblinear) considering L2, L1 and no penalty term and three possible C values (0.001,0.1,1.0)[9]. In the binary case, the best result was achieved in the Complete case with a newton-cg solver with L2 penalty and C=1.0. For the multi-class scenario, N-grams work slightly better, with a lbfgs solver with no penalty and C=1.0.

[5]Available at https://github.com/VinAIResearch/BERTweet (21/01/23).

[6]More information about NLTK library is specified in (Bird et al., 2009)

[7]Extracted using as reference all-MiniLM-L6-v2 HuggingFace sentence transformer model. More info available at https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (11/02/23).

[8]Results are in the form: Accuracy/F1±standard deviation.

[9]In case of incompatibility, we simply skipped that parameter configuration

With RF, we tested forests with gini and entropy split criteria, 2,4,6,8,10 as minimum samples in a node to split (min_samples_split) and 1,2,3,5,8,10 as minimum samples in a node to be considered a leaf (min_samples_leaf). In the binary case, the Embedding has best performances with gini index, 8 as min_samples_split and 3 as min_samples_leaf. For multi-class, N-grams has best performances with the same splitting criterion, 4 as min_samples_split and 1 as min_samples_leaf.

With MLP, for output layers, logistic sigmoid function was used in the binary context and softmax in the multi-label ones, testing relu, logistic sigmoid and tanh as functions for the hidden layers with lbfgs, sgd and adam as possible weight update algorithms with a 1e-4 tolerance for convergence and L2 regularization with λ=0.0001, keeping the other scikit-learn default parameters, testing (8,16,32), (16,32,64,128) and (64,128,256) as possible hidden layers, adopting an early stopping strategy considering in model assessment 10% of the training data as validation to perform the stop in case of necessity. This last solver resulted to be the best in the Complete binary classification case, with adam solving algorithm and logistic activation functions. In the multi-class case, tanh function and lbfgs solver, with a default (64, 128, 256) as network topology resulted to be the optimal choice and the overall best result.

3.2 BERTweet model

This section is dedicated to the assessment of BERT transformers technique for misogyny identification tasks: we will focus on differences in terms of performance with previously analyzed classifiers, specifically taking into account a BERT-based architecture model pre-trained on tweets according to the RoBERTa pre-training procedure[10]. We follow a similar pipeline with respect to the previous classifiers, applying TweetNormalizer paired with a specific tokenizer for BERTweet models employed in order to render the input suitable for HuggingFace's transformers[11]. For model selection, we performed a stratified cross-validation with k = 5 with the same folds defined for previous solvers and, setting as possible configurations 5 total epochs, with a train and validation batch of 16 and 8 respectively, 500 as warmup steps and 0.01 or 0.0001 as weight decay parameter, testing two different learning rates (1e5, 3e-5). Since in cross-validation for both the binary and multi-class problems the same model (having 0.0001 as weight decay and 3e-05 as learning rate) resulted from the selection process, we summarize test results for this specific model for both tasks in Tables 6 and 7.

Labels	Precision	Recall	F1-Score
0	74%	74%	74%
1	70%	70%	70%

Table 6: BERTweet binary task. Labels 0 and 1 stand for 'not misogynous' and 'misogynous', respectively

Labels	Precision	Recall	F1-Score
derailing	25%	18%	21%
discredit	60%	86%	71%
dominance	81%	0.49%	61%
sexual_harassment	57%	70%	63%
stereotype	92%	80%	85%

Table 7: BERTweet multi-class task

Acceptable results are obtained for the binary task, less desirable ones for the multi class case where we end up with 0.60 as F1 macro averaged score. In order to tackle the issue, we perform a further pre-processing on the data-set used for the multi-class task: for the sake of completeness, we perform a model selection considering MLP in the entire feature space, the best standard classifier, and BERTweet starting from a data-set expanded using Easy Data Augmentation (EDA) techniques as proposed in (Wei and Zou, 2019), artificially increasing the number of tweets: for each of the four under-represented categories, we added 50% more tweets for that particular category generating for each record new instances using Synonym Replacement (replacing 20% of the words in the tweet) and Random Swap (also considering 20% of words). As Table 8 shows, macro average score slightly increases in both cases, with better results for BERTweet in the model selection context[12].

Metric	Multi-Layer Perceptron		BERTweet	
	Mean VL	TS	Mean VL	TS
Accuracy	37.49% ± 1.91%	46.13%	75.71% ± 2.61%	45.82%
F1-score	62.56% ± 1.07%	50.28%	70.80% ± 2.65%	61.00%

Table 8: BERTweet and MLP comparison

Even though overall results tend to improve, we could still aim for higher performances given

[10]Consider again (Nguyen et al., 2020).

[11]More detailed information about the library is reported in (Wolf et al., 2019)

[12]We specify here that the BERTweet extracted from this model selection had a learning rate of 3e-05 with 0.01 as weight decay parameter, while the MLP kept the parameters of the best model found in the non augmented data-set.

the objective of using the model on new tweets. This desideratum, combined with the very poor frequency of *derailing* labels and the conceptual similarity between *dominance* and *stereotype* tweets, leads us to the choice of removing from the analysis the least common class and merge the other under-represented labels together, simplifying the task to a 3 total labels multi-classification problem. Figure 4 show the new distribution of the tweets with respect to the categories of misogyny they express.

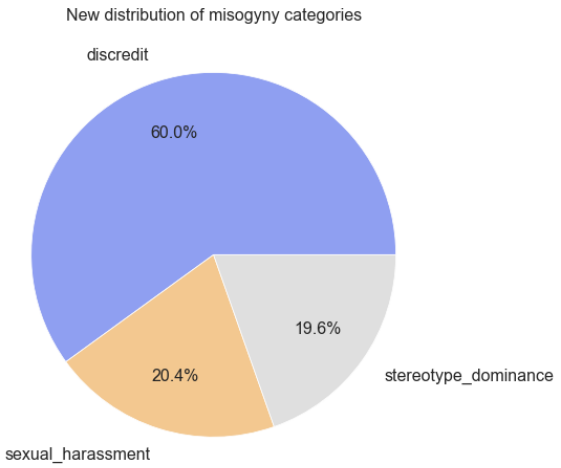

Figure 4: Distribution of tweets with respect to 'misogyny_category' after the reduction of their number

Hence, an additional model selection and assessment phase is performed taking into account only BERTweet classifiers with respect to a data-set with 3 labels, giving us the final model to apply on newly extracted tweets[13]. Results on test set of this final model are reported in Table 9, with 73% overall accuracy and 70% as macro-averaged F1 score.

Labels	Precision	Recall	F1-Score
discredit	60%	84%	70%
sexual_harassment	57%	70%	63%
stereotype_dominance	91%	68%	78%

Table 9: BERTweet multi-class task

4 Analysis of misogyny across occupations

Given the previously defined best model and the new tweet data-set manually extracted, we proceed with the main contribution of this report: understanding the relationship between a woman occupation and the kind of misogyny she is victim of on social media. To do so, we automatically classify a total of 760 new misogynous tweets extracted from Twitter using as keywords the most frequent hate-speech related words in the original data-set, focusing on women belonging to distinct occupation fields: in order to perform a naïve estimate of the trained classifier's generalization capabilities on the new tweet data-set, a sub-set of 187 tweets has been manually labeled with the three classes for comparison with the predicted outputs given by BERTweet: 0.65 of accuracy was reached by the model with respect to our gold standard, with 0.59 as macro-averaged F1 score.

4.1 Misogyny distribution across occupations

We start with a simple bar-chart depicting how different kinds of occupations might be related to distinct types of misogynous speech:

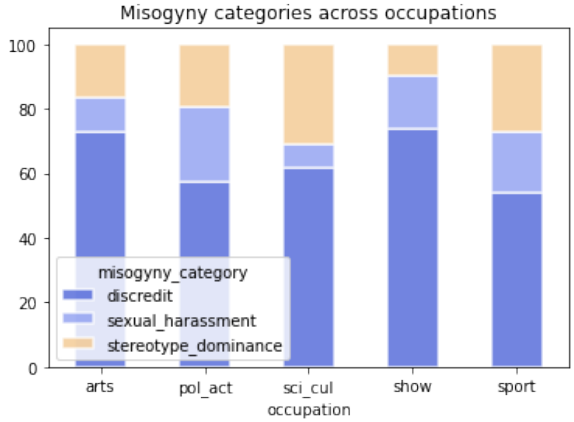

Figure 5: Different kinds of misogyny across occupations

The sci_cul category, followed by the sport one, seems to include a prevalence of stereotype_dominance and discredit tweets more then other categories: this could be a reasonable result considering how those kinds of occupations could be labeled as more "masculine", so that misogynous tweets might revolve around attacks regarding women stereotypes. Another reasonable result is related to the arts category, which is in great majority comprised of discredit tweets, while pol_act has the highest percentage of sexual_harassment kind of tweets.

4.2 *K*-medoids clustering analysis

We check if a simple clustering approach such as *K*-medoids is capable of making reasonable distinctions among various kinds of misogyny and among

[13]The last best model parameters result to be the same of the best one in the previous cases.

different kinds of occupations on the automatically labeled data-set, testing with different Ks and reporting the best identified result with K=9: in order to properly perform the clusterization, TweetNormalized texts have been transformed into vectorial representations using the same approach for the standard classification task in the embeddings context, choosing cosine similarity as distance measure for the algorithm. Results are shown in Figure 6 and 7.

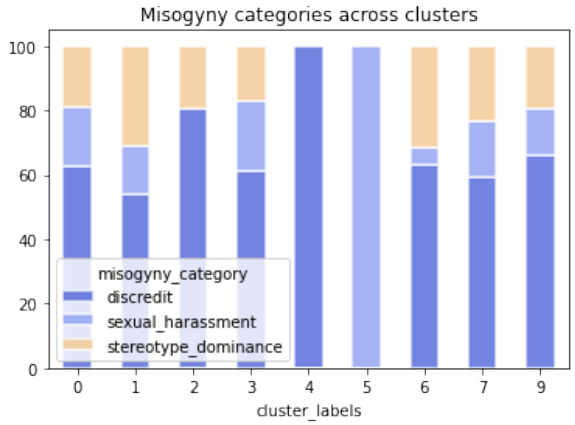

Figure 6: Different kinds of misogyny across clusters

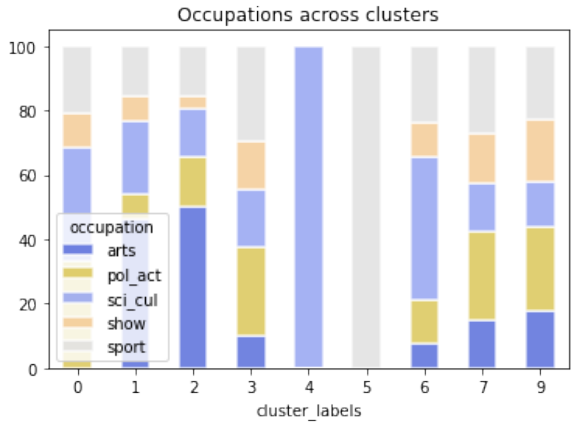

Figure 7: Different kinds of occupations across clusters

The approach does not lead to satisfactory results, even if something can still be said about clusters 4 and 5, respectively capturing two wholly distinct kind of occupations and kind of misogyny: following the clustering results, we would claim that women belonging to sci_cul category seems to be mainly victim of discredit, whereas sexual harassment is more present for women related to sports. To be complete, we emphasize the fact that medoids chosen as the two final clusters centers are not really intuitively prototypical for their respective classes: for tweet 4, the most interesting lemmas included in the medoid are "artist", "whore", "prostitute", "petty" and "vain", while for cluster 5 we have "men", "fight", "feminism".

4.3 Topic modeling

The second analysis on the data-set of misogynous tweets coming from both the original data-set and our newly extracted data was aimed at identifying some relevant topics within tweets according to the professional field they belong to. Relying on the functionalities provided by the Gensim library[14], we extracted three topics per occupation, taking into account not only unigrams, but also bigrams and trigrams.

Figures 8-12 show the representation of the first topic of each group in the form of a word cloud. It can immediately be seen that the most salient terms are usually offensive for all five classes, but are not explicitly related to professions in any way. Although this could mean that topic modeling does not bring useful contributions to our research, it is something that we expected. Indeed, as explained in Section 2.2, the selection of the new tweets was based on a number of very common words within the original data-set and these are exactly the ones that appear in these figures.

However, if we look at the less relevant terms, those that are smaller in the clouds, we notice that some topics contain expressions that are closely related to the semantic area to which the occupations belong. The most evident case regards the 'arts' category (8), that exhibits six references to disciplines pertaining to the world of arts ('design', 'photographer', 'art', 'photograph', 'designer', 'artist'). Such kind of references can also be found in the topics of classes 'sport' ('sport' and 'play') and 'sci_cul' ('space'). The opposite extreme seems to be represented by the topic concerning the category 'show', that apparently do not include any distinctive term. Nevertheless, by a deeper analysis it can be noted that it has a higher number of expressions ('fat', 'ugly', 'big', 'look', 'ass') that refer to the physical appearance of the victims. This evidence is perfectly consistent with the fact that the appearance of a woman has a pivotal importance in the show business, thus becoming the main target of haters' attacks.

[14]https://radimrehurek.com/gensim/ (07/02/23).

Figure 8: First topic extracted for the class 'arts'

Figure 9: First topic extracted for the class 'sport'

Figure 10: First topic extracted for the class 'sci_cul'

Figure 11: First topic extracted for the class 'pol_act'

Figure 12: First topic extracted for the class 'show'

4.4 Degree of negativity analysis

Our last original proposal consists in the assessment of each tweet's degree of negativity/aggressiveness exploiting an external *lexicon* built to estimate the level of valence, dominance and arousal of approximately 20000 English words[15]. To our current knowledge, there are no previously defined *lexica* specifically designed for the evaluation of the negativity/aggressiveness of a word. Hence, we decided to estimate such level following the intuition that words having high arousal values and low valence should be considered quite aggressive, while high valence and high arousal terms should be identified as positive/non-aggressive. Therefore, the negativity of a word is identified as the ratio between its arousal value over its valence value: the higher the ratio, the more negative a word should be considered. The intuition is supported observing that words a speaker would probably evaluate as quite aggressive tend to have a very high arousal-to-valence-ratio, while words more frequently associated with positive emotions present a lower ratio.

In order to assign a degree of negativity to a word *w* occurring in a tweet but absent in our chosen *lexicon*, we look for the closest word to *w* in a vectorial semantic space defined considering 2 billions tweets using the GloVe unsupervised learning algorithm[16] and assigning to *w* the negativity degree of such closest word, i.e. the *w'* in the *lexicon* having the highest cosine similarity score when compared to *w*. The "reliability" of the assignment is directly identified with the similarity score used to perform the assignment: a low cosine similarity between *w* and *w'* should make us less prone to accept the fact that *w* has about the same negativity score as *w'*. On the other hand, a high cosine similarity makes the assignment more reliable. Then we simply identify the negativity/aggressiveness of a tweet with the mean negative scores of its components, after having performed proper adjustments to the tweet[17]. We also identify the "reliability" of the negativity evaluation of a tweet as the mean reliability of each word's score. Table 10 illustrates relevant lemmas belonging to examples of most negative tweets in each occupation category, considering only tweets with a "reliability" higher then 0.7. We report interesting examples concering lemmas present in the most negative tweets along with their aggressiveness score.

[15] Further information are reported in (Mohammad, 2018)

[16] More information are reported in (Pennington et al., 2014)

[17] Removing stop-words, urls and @-mentions and having performed a lemmatization of the words that comprise the tweet

Occupation	Lemmas	Negativity	Reliability
arts	fucking, bitch, rot, hell	12.55	1.00
sport	victim, mentality, totally, love, rape	3.79	1.00
sci_cul	demand, bitch, moral, stand, shit, be, not, kid	9.41	0.98
pol_act	instagram, remove, comment, whore, guideline, smiling_halo	8.86	0.81
show	awful, woman, complete, skank, disgust, marked, filthy, idiot	4.94	1.00

Table 10: Examples of relevant lemmas in negative tweets for occupation category

We also report that the total mean negativity score for each occupation field, evaluated as the mean negativity score obtained by previously isolated tweets, is similar among each class, with sci_cul, arts and sports respectively having a score of 3.35, 3.40, 3.04 and with pol_act and show having scores of 2.48 and 2.44. We could say that the main negativity of the identified tweets depends upon words which are aggressive *per se* and shared among tweets belonging to various classes so that the kind of insulting words used in misogynous tweets tend to be the same across occupation fields and what relevantly changes are extra lemmas capturing aspects more related to the social position or work in question[18].

5 Conclusions

In order to summarize previous sections' contents, we state that results related to our attempted contributions on the misogyny identification problem on social network could be declined as follows: we focused on a comparison between different classifiers on the same two misogyny identification tasks. BERTweet ended up being the best classification model when compared with more "classical" solvers both for the binary and multi-class tasks, especially when paired with NLP augmentation techniques; with this model, we performed automatic labeling of newly extracted tweets taking into account different occupations; results for this section showed how different kinds of hate speech towards women is declined differently across their social roles. We then focused on the chance of identifying the negativity of misogynous tweet taking as a reference a *lexicon* characterizing words according to arousal, dominance and valence properties and we arrived at the conclusion that different occupations seem to share the same degree of negativity, defined by some constant lemmas paired with topic specific terms.

The main future prospects concerning a possible extension of our analysis are related to a deeper check of the test set's correctness and possible corrections of non-misogynous tweets, too. Different state-of-the-arts classifiers could be tested and more accurate aggressiveness measures could be defined, capable of capturing distinct level of negativity across the data-set.

[18]For instance, consider the presence of "instagram" as lemma in one of the most negative tweets in pol_act or the presence of "moral" in sci_cul

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples*. Accademia University Press.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.