Users Hate Blondes: Detecting Sexism in User Comments on Online Romanian News

Andreea-Codrina Moldovan[1], **Karla Csürös**[2], **Ana-Maria Bucur**[1,2], **Loredana Bercuci**[2]
[1]Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania
[2]West University of Timișoara, Romania
moldovanandreeacodrina@gmail.com
{karla.csuros98, loredana.bercuci}@e-uvt.ro
ana-maria.bucur@drd.unibuc.ro

Abstract

Romania ranks almost last in Europe when it comes to gender equality in political representation, with about 10% fewer women in politics than the E.U. average. We proceed from the assumption that this underrepresentation is also influenced by the sexism and verbal abuse female politicians face in the public sphere, especially in online media. We propose a novel dataset with sexist comments in Romanian language from online newspaper articles about Romanian female politicians and experiment with baseline models using classical machine learning models and fine-tuned pre-trained transformer models for the classification of sexist language in the online medium.

1 Introduction

While considerable progress has been made to combat sexism in the domain of political power, we are still a long way from achieving gender balance, especially in Eastern European countries like Romania. According to the Gender Equality Index published by the European Institute for Gender Equality in 2021, in Romania, only 25.8% of ministers, 19.7% of members of parliament, and 18.4% of the members of general assemblies are women (Barbieri et al., 2021). Romania ranks third to last (above Hungary and Malta by a small margin) regarding gender balance in political representation. It is much lower than the European Union's average (30.7% ministers, 31.5% members of parliament, and 29.3% members of general assemblies). Furthermore, women in leadership positions face discrimination and social pressure to conform to gender roles.

In the public sphere, female politicians in Romania face verbal abuse even from their peers. This abuse becomes even more prominent online, where the so-called disinhibition effect leads to exaggerated behaviours, from increased self-disclosure to increased verbal violence (Joinson, 2007; Suler, 2004; Wright, 2014; Wright et al., 2018; Rodríguez-Sánchez et al., 2020). Because the online environment offers anonymity and therefore lack of consequences for violent language, a variety of sexual slurs and sexist language appear online. This is especially true in the comment section of popular online newspapers when the topic of the articles are prominent women. The present paper uses these comments to build a dataset of sexist and non-sexist texts and experiment with several baselines models to detect the sexist language in Romanian automatically. The automatic detection of sexism could aid efforts to filter sexist texts, to encourage the prevention of, sensitization to, and sanctioning of such language.

WARNING: This paper contains sexist language.

2 Related Work

Automated methods for sexism detection have been implemented using a wide range of approaches for several languages such as English Rodríguez-Sánchez et al. (2020), Chinese (Jiang et al., 2022), Spanish (Rodríguez-Sánchez et al., 2021), French (Chiril et al., 2020a) or Arabic (Zahir et al., 2020). To the best of our knowledge, there are no such studies involving Romanian. However, one Romanian dataset for offensive language has been released by Manolescu and Çöltekin (2021). Recent studies on sexist language have been included in a review conducted by Istaiteh et al. (2020), which also includes a literature review of racist language.

Rodríguez-Sánchez et al. (2020) explore the more or less subtle ways in which sexist language manifests on Twitter in English and Spanish. The authors use a series of classical machine learning algorithms (Logistic Regression, Support Vector Machine and Random Forest) along with a Bidirectional Long Short-Term Memory model. Jha and Mamidi (2017) discuss the ambivalence of sexism

and how it can be both hostile and benevolent, and achieve the best results using a FastText classifier for distinguishing between the different kinds of sexism.

There are other similar approaches analysing how sexism manifests in different environments: at the workplace (Grosz and Céspedes, 2020), in the gaming communities (Ghosh, 2021), in politics (Gorrell et al., 2020; Fuchs and Schäfer, 2020). Several shared tasks are dedicated to the identification of online sexism and misogyny: sEXism Identification in Social neTworks (EXIST) (Rodríguez-Sánchez et al., 2021), Automatic Misogyny Identification at IberEval (Fersini et al., 2018) and Evalita (Fersini et al., 2020) and Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022).

This paper proposes a novel dataset for sexist language identification in Romanian language by collecting and annotating a corpus of online comments from newspaper articles about Romanian female politicians. Furthermore, we perform experiments on this corpus using classical machine learning models and fine-tuned pre-trained transformer models, thus providing baselines for comparison in future work.

3 ROFEMPOL Corpus

This data used in this study is part of the ROFEMPOL Corpus [1], which was compiled from online Romanian-language comments about Romanian female politicians. Ten of the most prominent female Romanian politicians were chosen: Clotilde Armand (USR[2]), Viorica Dăncilă (PSD[3]), Gabriela Firea (PSD), Monica Anisie (PNL[4]), Maria Grapini (PSD), Elena Udrea (PDL[5]), Diana Șoșoacă (AUR[6]), Carmen Dan (PSD), Lia Olguța Vasilescu (PSD). We have also included the prosecutor Laura Codruța Kovesi (DNA[7] / EPPO[8]) as, although she is not technically a politician, she was an influential figure in the Romanian political scene.

[1]the corpus will be made available upon request by contacting the authors of this paper

[2]Uniunea Salvați România (EN: Save Romania Union)

[3]Partidul Social Democrat (EN: Social Democratic Party)

[4]Partidul Național Liberal (EN: National Liberal Party)

[5]Partidul Democrat Liberal (EN: Democratic Liberal Party)

[6]Partidul Alianța pentru Unirea Românilor (EN: Alliance for the Union of Romanians)

[7]Direcția Națională Anticorupție (EN: National Anticorruption Directorate)

[8]European Public Prosecutor's Office

3.1 Data Collection

A sample of 2022 comments about the aforementioned female politicians was extracted from online comments sections. All comments were available online, either from the comments section of online newspapers or Facebook pages. We have selected the most popular and active Romanian online newspapers and extracted the comments from the public news articles. The comments from Facebook were selected from the public profiles of female politicians and the online newspapers' public pages.

The comments were manually extracted between November 2020 and January 2021 and reflect the salient political issues of the time. Two raters annotated the corpus, both female researchers in the field of Gender Studies, who evaluated each comment as sexist (1) or non-sexist (0).

3.2 Annotation Criteria

There are many studies that classify sexist speech and offer annotation criteria in English (Frenda et al., 2019; Parikh et al., 2019; Southern and Harmer, 2019), French (Chiril et al., 2020b) and Indian languages (Bhattacharya et al., 2020). Handrabura and Gherasim (2018) wrote one of the most in-depth practical guides for non-sexist speech in Romanian, classifying some of the most common types of sexist language in Romanian. As such, our annotation criteria for sexist/non-sexist texts are partially based on these studies and include: a) Gendered Violence; b) Gendered Insults; c) Role Stereotypes; d) Female Titles; e) Sexualisation and Physical Appearance.

a) Gendered Violence

By gendered violence, we mean language-based sexual harassment, inciting sexual harassment and threats of physical abuse, rape, or murder. Our understanding of gendered violence is based on the concept of "cyberviolence against women and girls" (cyber VAWG), which is defined as the "full spectrum of behaviour ranges from online harassment to the desire to inflict physical harm including sexual assaults, murders and suicides" [9]. The 2015 report includes harassment in its categorization of cyber VAWG, defining it as "the use of technology to continuously contact, annoy, threaten, and/or scare the victim". While online comments are not sent directly to the victims, they act as defamatory

[9]https://www.broadbandcommission.org/publication/cyber-violence-against-women/ (last accessed April 10, 2022)

language, which affects the target. We, therefore, considered these comments as harassment.

b) Gendered Insults

Online harassment often contains gendered insults, related especially to slut-shaming or shaming for nonconformity to societal expectations. Gendered insults represent any words or phrases used disproportionately against a specific gender and are linked to the perpetuation of stereotypical societal beliefs about that particular gender, in this case, women. Gendered insults can be i) Sexual or ii) Non-sexual.

Table 1: Samples from the ROFEMPOL corpus containing sexual insults

NU A FOST IUBITA? Daca a fost	***prostituata***	normal!
EN: WASN'T SHE LOVED? If she was a	***prostitute***	it's normal!
Are dreptate	***fufa***	asta.
EN: This	***philanderer***	is right.
iti rup gatul instant...	***curva***	tradatoare de neam si tara
EN: I'll break your neck instantly... you	***whore***	nation and country traitor
E iubita... A fost Basescu's	***BITCH***	O ordinara imputita...
EN: She is the lover... She was Basescu's	***BITCH***	A stinking lowlife...
Vad ca pe o	***tarfa***	pesedista o cam deranjeaza
EN: I can see that a	***whore***	of PSD is a little bothered
Da,	***Matracuca***	lui Pandele
EN: Yes,	***the bimbo***	of Pandele
Iar tu vei face iar pușcărie,	***zdreanțo***	
EN: You will go to jail,	***slut***	

i) Sexual

The gendered insults in this category denote the users' beliefs that female politicians are sexually promiscuous and have engaged in sexual favors to advance their political careers. Examples include: "prostituată" (EN: "prostitute"), "fufă" (EN: "philanderer"), "curvă" (EN: "whore"), "BITCH", "tarfă" (EN: "slut"), "matracucă" (EN: "lower class and unintelligent woman"), "zdreanță" (EN: "disgraced woman" and "rag"). We present some examples in Table 1.

ii) Non-sexual

In this category, we have aggregated insults that do not discuss the politicians' sexual lives, but their personalities, attitudes and social statuses. Notably, in Romanian, the following examples are solely used with reference to women: "tută" (EN: "dumbass"), "țață" (EN: "vulgar woman"), "mahalagioacă" (EN: "loud lower-class woman who lives in the ghetto"), "divă", "țoapă" (EN: "boor").

c) Role Stereotypes

Related to the issue of gendered insults is that of role stereotypes. A gender stereotype is a preconception about the characteristics and societal roles that gender should or should not have. Pickering (2001) argues that stereotypes create "difference as deviant for the sake of normative gain". As such, gender stereotypes work to uphold patriarchal structures, and, as others (Chiril et al., 2020b) have argued, they perpetuate gender-normative attitudes. For our corpus, we divided female stereotypes into i) "Women's jobs" and ii) "Female attitudes."

Table 2: Samples from the ROFEMPOL corpus containing stereotypes portrayed as "Women's jobs"

Sa stea	***acasa***	sa. Si creasca copiii
EN: She should stay	**at home**	to. And to raise the children
minte nici cât o bibilică!!! Stai	***acasă***	și croșetează nu te mai face de rahat
EN: birdbrain!!! Stay	***home***	and knit, don't embarrass yourself
acum ca e	***casnica***	la rosiorii de vede
EN: now that she is a	***housewife***	at rosiorii de vede
Eroină, dar nu și	***mamă***	... :-(
EN: A hero, but not a	***mother***	... :-(
Acum totul e o nebunie... tu esti	***mamă***	soție lasa i naibii de politicieni si vezi
EN: Now everything is insane... you are a	***mother***	a wife, forget those goddamn politicians
dar sunt sigur ca si in	***bucatarie***	le mai incurci.
EN: but I am sure that in	***the kitchen***	you also mess up sometimes.
Apucă-te Vasilico de	***sarmalele***	alea, lasă basmele!
EN: Vasilico, start cooking those	***cabbage rolls***	, forget the fairytales!
O fi terminat de făcut	***zacusca***	și acum se plictisește?!
EN: Maybe she finished making the	***zacusca***	and now she is bored?!
doamne ce moaca de	***bucatareasa***	nepregatita are, un jeg, un gunoi
EN: god, she looks like a unskilled	***cook***	, dirtbag, she is garbage
	Spalatoreasa	asta il ironineaza pe Klaus
EN: This	**washerwomen**	mocks Klaus
grapini zici ca e	***femeia de servici***	acolo, are grija sa nu se faca mizerie
EN: grapini looks like	***the cleaning lady***	there, she takes care not to make a mess
o rapandula	***secretara***	din videle este adevarat- ce spune
EN: A slut,	***the secretary***	from videle, it's true what they say

i) "Women's Jobs"

"Women's jobs" can be defined as being related to domesticity, namely being a wife, a mother, or a housekeeper. Despite their accomplishments, female politicians are oftentimes ridiculed for not having children. Users also make attempts to dismiss female politicians by claiming that they are impostors who, in fact, have stereotypically female careers (e.g., cook, washerwoman, cleaning lady, secretary) (Table 2).

ii) "Female Attitudes"

On the one hand, multiple comments employ positive stereotypes to talk about female politicians' success, attributing it to them being caring, maternal, soft or submissive. However, on the other hand, we also have examples of negative stereotypes, such as women being portrayed as hysterical

("isterică"), angry ("nervoasă", "crizată"), gossipy ("bârfitoare") as seen in Table 3.

Table 3: Samples from the ROFEMPOL corpus containing stereotypes portrayed as "Female attitudes"

Eu tot o spun, o	***femeie***	nu are un rol activ
EN: I keep saying it, a	***woman***	doesn't have an active role
sa fie condusă de femeie, numai o	***femeie***	poate avea grija de cetățenii ei precum
EN: to be led by a woman, only a	***woman***	can take care of the citizens like
EA insasi fiinta, parte din feminin,	***matern***	umanitate, parte din istoria
EN: SHE is being itself, part of the feminine,	***maternal***	humanity, part of history
USRist	***isterică***	prinsă cu mâța-n sacul de voturi?
EN: The USRite is	***hysterical***	when caught with her hand in the vote jar?
Femeia este precum o piranda	***nervoasa***	dispusa oricand sa-si ridice
EN: The woman is like a gypsy who is	***angry***	ready at any time to draw up
trebuie sa muncească nu sa	***bârfească***	Firea a fost numită
EN: she must work, not	***gossip***	Firea was named
D-na Grapini era singura și cea mai	***crizată***	doamnă din sală. Mă bucur că a
EN: D-na Grapini was the only and most	***hysterical***	lady in the room.

Table 4: Samples from the ROFEMPOL corpus containing mocking female titles

EN:	***Cucoana*** ***Lady***	cplm te recomanda sa fii primar what the hell recommends you to be mayor
EN:	***Tanti*** ***Auntie***	Nuti,daca ai spune adevarul, tu si Nuti, if you told the truth
de unde are atâta tupeu această	***madamă***	
EN: where does this	***madame***	get the nerve
ce Marete Realizari a inregistrat	***DUamna***	Gabi? Cum a inceput Macar
EN: what Great Achievements doe	***LAdy***	Gabi have? How did she start
Cui ii mai pasa de	***domnisoara***	batrana? Sa stea acolo fara
EN: Who cares about this	***spinster***	? She should stay there without
Bravo	***fata***	sper sa fii ca o pumă
EN: Well done	***girl***	I hope you will be like a puma
nu inteleg cine a pus-o pe aceasta	***fetiscana***	sa candideze ca primar
EN: I don't understand who made this	***schoolgirl***	run for mayor
Viata va fi foarte grea pentru	***ex-ministruoasa***	si cei 3-4 iubiti
EN: Life is very hard for the	***ex-ministress***	and the 3-4 lovers
Ce aveti cu biata	***ministreuză***	Ea a răspuns correct.
EN: What do you have against the poor	***ministreaseuse***	She answered corrrectly.
Doamna	***primăriță***	tace și face! Si-a inceput mandatul
EN: Madam	***mayoress***	gets the job done!

d) Female Titles

Users generally employ mocking female titles when addressing female politicians. Examples include: "cucoană" (EN: lady), "tanti" (EN: auntie, referring to an older woman), "madamă" (adapted from the French madame), "duamnă" from "doamnă" (EN: Mrs.), "domnișoară" (EN: Miss) (Table 4). Young or young-looking politicians are also regularly addressed as "fată" or "fetișcană" (EN: girl or girlie). There are also cases of users creating female versions of male titles with negative connotations: e.g., for the female minister, "ministru" (EN: minister) is mixed with "monstruasă" (EN: monstrous) to create "ministruoasă" or with "stripteuză" (EN: stripper) to create "ministreuză". Another example is "primăriță" (EN: female mayor), which adds the diminutive "-iță" to "primar" (EN: mayor). Such forms of address are used ironically in an attempt to deride the politicians and to express contempt.

e) Sexualisation and Physical Appearance

Comments on the physical appearance of the female politicians were very frequent, even though they were generally irrelevant to the topics of the news stories. Physical appearance comments include comments on: i) Body Weight, ii) Hair Color Stereotypes, iii) Clothes and Style, and iv) Perceived Physical Attractiveness.

i) Body Weight

Body shaming comments were more frequent in the case of female politicians who do not conform to the "thin" beauty standard. In this case, politicians were directly called "grasă" (EN: "fat"), or were compared to animals and fantastical creatures that are stereotyped as being overweight, e.g., "porc" (EN: "pig") or "matahală" (EN: "bugbear") as seen in Table 5.

Table 5: Samples from the ROFEMPOL corpus containing body shaming

Ca vrea cineva sao imbolnaveasca pe	***grasa***	asta? E cantitate nesemnificativa.
EN: Who wants to get this	***fatso***	sick? She doesn't matter.
un	***porc***	de o asemenea greutate
EN: a	***pig***	of this weight
in	***matahala***	asta de Sosoaca isi gasesc modelul
EN: they see this	***hulk***	Sosoaca as a model

Table 6: Samples from the ROFEMPOL corpus containing hair color stereotypes

	Blondo	vezi ca ieșim in strada
EN:	***Blondie***	we will riot
Năpârcă		ce vrea să pară inocentă(!)
	blondă	
EN: A	***blonde***	adder who wants to look innocent(!)

ii) Hair Color Stereotypes

The most common stereotype about hair color in the corpus is the "dumb blonde". As such, blonde female politicians received a significant number of comments in which users would make derogatory references to their hair color, often identifying them with it (Table 6).

iii) Clothes and Style

News stories that feature full-body pictures of female politicians often drew comments about their

fashion style and choices. For example, references were made to the cleanliness of their clothes, the shortness of their skirts ("fustă scurtă"), as well as the expensiveness of their outfits (i.e., clothes and accessories), as seen in Table 7.

Table 7: Samples from the ROFEMPOL corpus containing sexists comments related to clothes and style

Mana Iute a imbracat in campanie o EN: Sticky Fingers wore, during the campaign, a	***rochie*** ***dress***	murdara tesuta cu motive romanesti.. that was dirty and with traditional motifs
păi puteai să stai EN: you could sit in a short	***fustă*** ***skirt***	scurtă și cu fundul pe birou așa and with your ass on the desk
valuta neagra spalata in tara, EN: foreign currency laundered in the country,	***haine*** ***clothes***	si accesorii de sute de mii de euro and accessories worth thousands of euros

Table 8: Samples from the ROFEMPOL corpus containing sexists comments related to perceived physical attractiveness

mai vedem aceasta agramata EN: can we still see this illiterate woman? an	***urata*** ***ugly***	femeie woman
 EN:	***Hidoșenia*** ***Hideousness***	trebuie mascată cumva... must be hidden somehow
dumneata esti, si ai fost o papusa EN: you are, and have always been a	***frumoasa*** ***beautiful***	:) si chiar: inca foarte frumoasa :)! doll :) and even: still very beautiful :)!
NU E EN: she IS NOT A	***FEMEIE*** ***WOMAN***	De-aia au si ales-o deviatii de la That's why the deviants have chosen her

Table 9: Samples from the ROFEMPOL corpus containing dehumanisation language

Baaa, v-ati uitat bine la EN: Maaan, have you looked carefully at this	***creatura*** ***creature***	asta? Dupa cum se imbraca ? Judging by the way she dresses
Bai EN: You brainless	***vaca*** ***cow***	descreierata, tu neaparat ai nevoie , you definitely need
La puscarie cu tine, EN: Rot in jail,	***javra*** ***bitch***	muista ordinara! disgusting cocksucker!
cu gândul la... os. Este ca o EN: always thinks of... boning. She is like a	***cățea*** ***bitch***	în călduri in heat
Sau ție îți place gunoiul, EN: Or you like garbage, you foreign	***scroafă*** ***sow***	alogenă! !

iv) Perceived Physical Attractiveness

Commenters offer judgements on how attractive or unattractive they find certain female politicians. Both negative – e.g. "urâtă" (EN: "ugly"), "hidoșenie" (EN: "monstrosity") – as well as positive – e.g. "frumoasă" (EN: "beautiful") – were included as being sexist, as they are both inappropriate in context. Moreover, for the "most unattractive" politicians, commenters questioned their female gender, claiming that they are not women ("nu e femeie") as seen in Table 8.

v) Dehumanisation

Dehumanisation occurs primarily through zoomorphism, i.e., by comparing the targets with female animals, often with sexual connotations. Some examples (Table 9) include "creatură" (EN: "creature"), "vacă" (EN: "cow"), "javră" (EN: "female dog", "bitch"), "scroafă" (EN: "sow").

After both raters annotated the data individually, we measured the inter-rater agreement using Cohen's Kappa. The coefficient can have values between -1 and 1, with a value equal to 1 meaning a perfect agreement between the annotators. The Cohen's Kappa coefficient for the annotations of the two raters in this paper is 0.87, meaning there was a good inter-rater agreement between the annotators. The two raters discussed the disagreements until a final common decision was made. The resulting ROFEMPOL dataset contains 1135 non-sexist samples and 887 sexist samples.

We explored the dataset by computing the keyness scores for the sexist and non-sexist texts (Kilgarriff, 2009; Gabrielatos, 2018). The keyness analysis is performed by comparing the frequencies of the words from the sexist comments (target corpus) to the frequencies of words from the non-sexist comments (reference corpus). In Figure 1 we report the top 15 words from the two classes ordered by their log-likelihood ratio (G^2) (Dunning, 1993).

The sexist comments contain some sexist keywords such as "madam" (EN: "madam"), "madame" (EN: "madam"), "mahalagioaică" (EN: "ghetto woman") or "coană" (EN: "lady"), while non-sexist texts contain more polite addressing forms such as "doamnă" (EN: "Mrs.") or "dumneavoastră" (EN: "you", the polite second person singular or plural in Romanian language).

4 Baseline Methods

In this section, we propose several baseline methods for the binary classification of sexist language using ROFEMPOL. We explore several encoding methods for the Romanian text data, such as Bag-of-Words and BERT-based sentence representa-

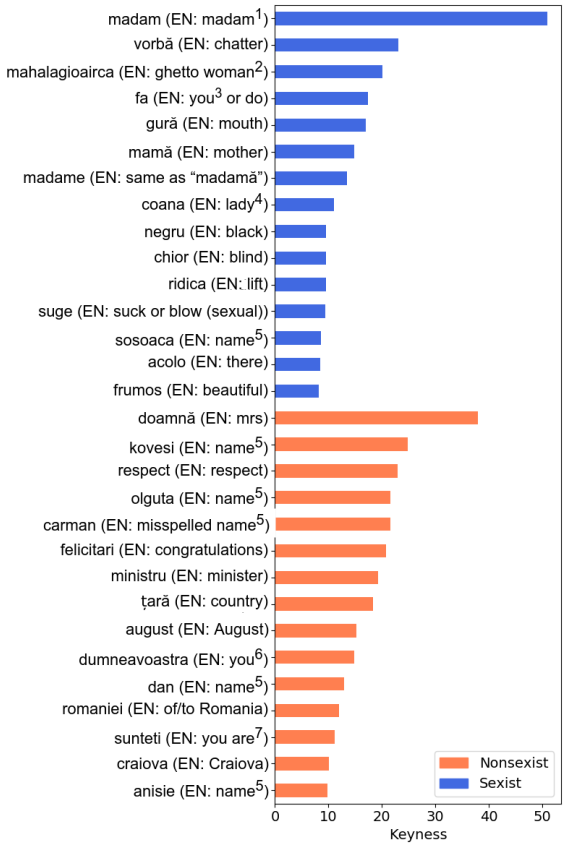

Figure 1: Keyness scores for the words from ROMFEM-POL corpus.

tions. Alongside classical machine learning models (i.e., Logistic Regression, SVM, Random Forest), we also explore fine-tuning pre-trained transformer models for the Romanian language.

4.1 Text Representation

Bag-of-Words (BOW) and Term Frequency–Inverse Document Frequency (TF-IDF) We chose to use BOW-based representations because they are language independent. In addition, BOW representations allow for modelling sexism based on keywords, as some sentences in the dataset can be easily identified based on certain sexist words (i.e., the keywords from Figure 1).

[1] sarcastic adaptation of the French madame; the word suggests the person does not deserve the title of lady, madam
[2] a loud lower class woman who is unrefined
[3] slang specifically used to address women directly, similar to lady, but it implies the woman is lower class
[4] it suggests that the woman addressed is older, unattractive and unrefined
[5] of one of the politicians
[6] the polite second person singular or plural
[7] the polite second person singular or plural, or plain form of second person plural

Multilingual BERT As opposed to the BOW and TF-IDF word representations, which do not contain any information about the context, sentence representations as given by modern transformer networks (Reimers and Gurevych, 2019) offer richer semantic information and have been successfully used in low-resource scenarios (Ranasinghe and Zampieri, 2021). As such, we use Sentence Transformer (Reimers and Gurevych, 2019) to extract embeddings from BERT-based models. We use a pre-trained Multilingual BERT (M-BERT) (Devlin et al., 2019) which was trained on 102 languages using Wikipedia text, including Romanian language.

Romanian BERT As opposed to M-BERT, the Romanian BERT (Ro-BERT) (Dumitrescu et al., 2020), is a more specialized model, trained on a larger Romanian corpus. Moreover, the tokenizer is better suited for handling Romanian texts, using fewer tokens to encode words than M-BERT, while also having fewer unknown tokens. The model was trained on a large corpus of Romanian data from Wikipedia, OPUS (parallel corpus with translated texts from the web) (Tiedemann, 2012) and OSCAR (Common Crawl data in Romanian language) (Suárez et al., 2019). We use Ro-BERT for extracting sentence representations.

4.2 Models

We evaluate the performance of classical machine learning classifiers: Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) on BOW and semantic sentence representations. Moreover, we directly fine-tune transformer models pre-trained on Romanian language text: M-BERT and Ro-BERT. The Ro-BERT model has been shown to outperform its multilingual counterpart, M-BERT, in downstream tasks such as named entity recognition and part-of-speech tagging (Dumitrescu et al., 2020).

5 Experiments and Results

In this section, we describe the classification experiments performed for the detection of sexist language in Romanian text and report the obtained results.

5.1 Experiments

ROFEMPOL Corpus is split into training and testing sets, with 1617 texts in the training split and 405 texts in the testing split.

Since the dataset contains a small number of

Table 10: Results for sexist language detection on ROFEMPOL. We report Precision, Recall and F_1 for each model on the two classes (Non-sexist and Sexist) and weighted averages. We also report Macro-F_1 score. The best performing model is the fine-tuned Ro-BERT.

	Non-sexist			**Sexist**			**Weighted Average**			
Model	Precision	Recall	F1	Precision	Recall	F_1	Precision	Recall	F_1	Macro-F_1
BOW + LR	0.68±0.01	0.78±0.02	0.72±0.01	0.66±0.02	0.52±0.02	0.58±0.01	0.67±0.01	0.67±0.01	0.66±0.01	0.66±0.01
BOW + RF	0.62±0.01	0.90±0.05	0.74±0.03	0.72±0.10	0.31±0.02	0.43±0.03	0.67±0.05	0.64±0.03	0.60±0.02	0.58±0.02
BOW + SVM	0.68±0.02	0.79±0.01	0.72±0.01	0.65±0.02	0.50±0.04	0.56±0.03	0.66±0.01	0.66±0.02	0.65±0.02	0.64±0.02
TFIDF + LR	0.70±0.01	0.79±0.01	0.74±0.00	0.68±0.00	0.56±0.03	0.61±0.02	0.69±0.00	0.69±0.01	0.68±0.01	0.68±0.01
TFIDF + RF	0.63±0.01	**0.90±0.04**	0.74±0.02	0.72±0.06	0.31±0.02	0.44±0.02	0.66±0.03	0.64±0.02	0.61±0.01	0.59±0.02
TFIDF + SVM	0.69±0.01	0.77±0.02	0.73±0.01	0.66±0.01	0.56±0.03	0.61±0.02	0.68±0.01	0.68±0.01	0.68±0.01	0.67±0.01
M-BERT emb + LR	0.68±0.01	0.80±0.01	0.74±0.00	0.67±0.01	0.53±0.01	0.59±0.01	0.68±0.01	0.68±0.01	0.67±0.01	0.66±0.01
M-BERT emb + RF	0.64±0.00	0.82±0.02	0.72±0.01	0.64±0.02	0.41±0.01	0.50±0.01	0.64±0.01	0.64±0.01	0.62±0.00	0.61±0.01
M-BERT emb + SVM	0.67±0.01	0.79±0.02	0.72±0.01	0.65±0.02	0.50±0.04	0.57±0.02	0.66±0.01	0.66±0.01	0.66±0.01	0.64±0.01
Ro-BERT emb + LR	0.70±0.01	0.78±0.00	0.74±0.01	0.67±0.01	0.57±0.03	0.62±0.01	0.69±0.01	0.69±0.01	0.69±0.01	0.68±0.01
Ro-BERT emb + RF	0.68±0.01	0.83±0.01	0.74±0.01	0.69±0.02	0.49±0.03	0.57±0.02	0.68±0.01	0.68±0.02	0.67±0.02	0.66±0.02
Ro-BERT emb + SVM	0.71±0.01	0.79±0.02	0.75±0.01	0.69±0.02	0.58±0.02	0.63±0.01	0.70±0.01	0.70±0.01	0.70±0.01	0.69±0.01
Fine-tuned M-BERT	**0.77±0.05**	0.75±0.02	0.76±0.02	0.67±0.06	0.70±0.03	0.69±0.02	0.74±0.02	0.73±0.01	0.73±0.01	0.72±0.01
Fine-tuned Ro-BERT	0.76±0.05	0.80±0.03	**0.78±0.01**	**0.76±0.06**	**0.71±0.03**	**0.73±0.02**	**0.77±0.01**	**0.76±0.01**	**0.76±0.01**	**0.75±0.01**

samples, we performed a 5-fold cross-validation for all models. For LR, RF and SVM, we perform a hyperparameter grid search on each fold to find the best hyperparameters for the models. The search space used for grid search for each model can be found in Appendix.

The pre-trained transformer models, Ro-BERT and M-BERT, are fine-tuned using the AdamW (Kingma and Ba, 2014) optimizer with a learning rate of 0.00001 with a linear decay with 50 warm-up steps. Due to computational limitations, we train the models for 4 epochs with a batch size of 4.

All the reported results are obtained from the 5-fold cross-validation. The performance of models from each fold is evaluated on the test split. We report the mean and standard deviation of the performance scores for the 5 splits for Precision, Recall and F_1-score, weighted averages and macro-F_1.

5.2 Results

The results of the classification experiments for sexism classification on the ROFEMPOL Corpus are presented in Table 10. The best performing model is Ro-BERT, the pre-trained BERT model for Romanian language, obtaining the overall best scores in discriminating between sexist and non-sexist texts. Ro-BERT attains a 0.75 Macro-F_1 score in the classification task, an improvement of 0.03 over M-BERT.

All the models perform better at classifying the non-sexist texts than identifying the sexist ones. The differences between the performances for the two classes are greater for the classical machine learning models using BOW, TF-IDF and BERT-based representations than for the fine-tunes models. The biggest gap in the performances between the two classes is found in the classification using Random Forest on Bag-of-Words representations, with 0.74 F_1-score for non-sexist and 0.44 F_1-score for the sexist class.

Comparing the three classical machine learning models, SVM and LR perform better than RF for all the text representation methods (BOW-based, BERT-based encodings), as seen in the Macro-F_1 score.

5.3 Qualitative Results and Discussion

We present a sample of correct and incorrect predictions from the best performing model, Ro-BERT, in Table 11 and 12.

Table 11: Selected correct predictions using Ro-BERT, the best performing model.

Non-sexist	**Sexist**
Asa ai timp de lectura. Când erai Primar nu era timp deloc	vrea si iea la festivalul homosexualilor in sibiu? jigodie pdlista vinzatore de tzara
EN: This way, you have time to read. When you were Mayor, you didn't have time at all	does she want to go to the homosexual festival in sibiu? PDL bitch, country traitor
un primar asa tot sa avem bravo ție olguta .un primar ambitios!	Cățeaua asta trebuie dusă acolo îi este locul - Jilava!
EN: we are glad to have such a mayor well done olguta .an ambitious mayor	This bitch must be taken where she belongs - Jilava!
Doamna Olguța trebuie sa-și sporească averea.	Veo, n-ai murături de pus?
EN: Mrs. Olguța has to increase her fortune.	Veo, don't you have pickles to make?

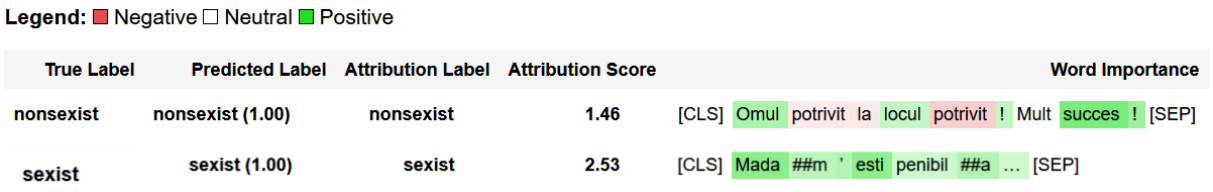

Figure 2: Visualizing salient tokens contributing to the prediction of the correct class using the Ro-BERT model. Term color indicates attribution intensity, red is negative, green is positive. EN: "The right man at the right place ! Good luck !", "Madam you are lame ..."

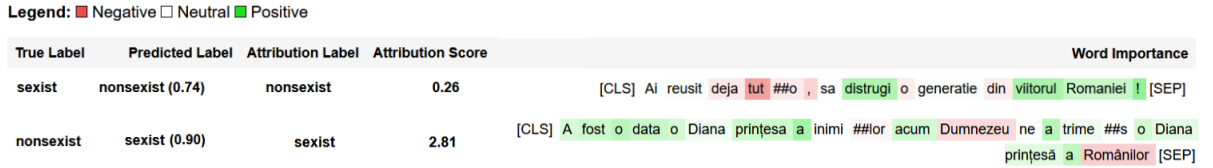

Figure 3: Visualizing salient tokens contributing to the prediction of the incorrect class using the Ro-BERT model. Term color indicates attribution intensity, red is negative, green is positive. EN: "Dumbass you have already managed to destroy a generation from Romania's future!", "Once upon a time there was a Diana princess of hearts, now God has sent us a Diana princess of the Romanians."

Table 12: Selected incorrect predictions using Ro-BERT.

Prediction: Non-sexist **True Label: Sexist**	**Prediction: Sexist** **True Label: Non-sexist**
Hudrea tot pe centura .... EN: Hudrea keeps working the streets	A fost o data o Diana prințesa a inimilor acum Dumnezeu ne a trimes o Diana prințesă a Românilor There once was a Diana, princess of hearts now God has sent us a Diana princess of the Romanians
Ai reusit deja tuto, sa distrugi o generatie din viitorul Romaniei!. EN: You have already managed, dumbass, to destroy a generation of the future of Romania!.	Clocosoros asta,e imaginea apocalipsei,intruchiparea raului,arata ca un paznic la poarta Infernului.... O meritati cretinilor ! This Clocosoros is the image of the apocalypse, the embodiment of evil, she looks like a guard at the gate of hell... You deserve her, idiots!
Felicitari pentru articol ....o TOAPA parvenita si needucata ! EN: Congratulations on the article... An uneducated boorish upstart!	Analfabetă, incultă , scursura societății ... You illiterate uneducated trash of society...

Regarding the incorrect decisions of the model, some of the texts predicted as being sexist contain offensive words, but the words are not sexist, according to the annotations guidelines, such as "analfabetă" (EN: "illiterate") or incultă (EN: "uneducated"). Other offensive words are not targeted towards female politicians but other people. For example, "cretinilor" (EN: "idiots") is targeting voters. The texts incorrectly labelled as non-sexist contain vulgar, sexist words such as "țoapă" (EN: "boor") and "tută" (EN: "dumbass") that can be easily recognised as being sexist. The Ro-BERT model may fail to recognise these words as sexist because it is not pre-trained on informal text found in the comments from news articles or social media.

Furthermore, we compute word importance (attribution scores) from Ro-BERT to interpret the model's predictions. We use Integrated Gradients (Sundararajan et al., 2017) from the Captum library (Kokhlikyan et al., 2020) to show the most salient tokens for the incorrect predictions. From the examples in Figure 2, we can conclude that the model can recognize "madam" being used as a sexist word in the samples from our corpus, thus having an important attribution in the decision of the model.

In Figure 3 we show two examples of word attributions in incorrectly predicted texts. In the first comment, even if the tokens *tut ##o* (EN: "dumbass") have strong negative attributions, the final decision is also influenced by the other tokens in the sentence, labelling the comment as being non-sexist. In the second text, the token "prințesă" (EN: "princess") has a strong positive attribution in the decision of the model to label the text as being sexist, although the word is not used as sexist in this context.

6 Conclusion

Our findings underline the fact that Romanian female politicians are relentlessly targeted and stereotyped because of their gender in the eye of the Romanian public, who uses sexist language to criticise them online. Automatic detection of sexist speech in Romanian language is a result of the imperative need to eliminate such forms of gender-based discrimination in order to work towards gender equality and a truly democratic Romanian society, one in which female politicians can thrive. We presented the novel ROFEMPOL dataset for sexist language identification in Romanian, collected from online comments from newspaper articles about Romanian female politicians. Furthermore, we performed experiments on this corpus using classical machine learning models and fine-tuned pre-trained transformer models, thus providing baselines for comparison in future work.

Further work on the ROFEMPOL corpus will attempt to include a larger dataset and to annotate the categories outlined in this paper. The ROFEMPOL dataset could also prove productive in a comparison with a corpus on male politicians, testing whether sexist speech is a general phenomenon for Romanian users or whether it is only targeted at female politicians. Lastly, ROFEMPOL could be included in a larger corpus on sexism in the Romanian language in an attempt to limit the spread of sexist language online.

References

Davide Barbieri, Antonio Garcia Cazorla, Laurène THIL, Blandine Mollard, Julia Ochmann, and Vytautas Peciukonis. 2021. Gender equality index 2021: Health.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression.

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. An annotated corpus for sexism detection in French tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Ted E Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. Ami@ evalita2020: Automatic misogyny identification. In *EVALITA*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Simona Frenda, Bilal Ghanem, Manuel Montes, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36:4743–4752.

Tamara Fuchs and Fabian Schäfer. 2020. Normalizing misogyny: hate speech and verbal abuse of female politicians on japanese twitter. *Japan Forum*, 33:1–27.

Costas Gabrielatos. 2018. Keyness analysis: Nature, metrics and techniques. In *Corpus Approaches to Discourse*, pages 225–258. Routledge.

Ayushi Ghosh. 2021. Analyzing toxicity in online gaming communities. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10):4448–4455.

Genevieve Gorrell, Mehmet Bakir, Ian Roberts, Mark Greenwood, and Kalina Bontcheva. 2020. Which politicians receive abuse? four factors illuminated in the uk general election 2019. *EPJ Data Science*, 9.

Dylan Grosz and Patricia Conde Céspedes. 2020. Automatic detection of sexist statements commonly used at the workplace. *CoRR*, abs/2007.04181.

L Handrabura and A Gherasim. 2018. Limbajul nonsexist: Repere conceptuale şi recomandări practice.

Othman Istaiteh, Razan Al-Omoush, and Sara Tedmori. 2020. Racist and sexist hate speech detection: Literature review. In *2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 95–99.

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, Vancouver, Canada. Association for Computational Linguistics.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Adam Joinson. 2007. Causes and implications of disinhibited behavior on the internet. *Psychology and the internet: intrapersonal, interpersonal, and transpersonal implications*.

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Mihai Manolescu and Çağrı Çöltekin. 2021. ROFF - a Romanian Twitter dataset for offensive language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 895–900, Held Online. INCOMA Ltd.

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.

Michael Pickering. 2001. *Stereotyping: the politics of representation*. Red Globe Press.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–13.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Rosalynd Southern and Emily Harmer. 2019. Othering political women: Online misogyny, racism and ableism towards women in public life. In *Online Othering*, pages 187–210. Springer.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 7:321–6.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Michelle Wright. 2014. Predictors of anonymous cyber aggression: The role of adolescents' beliefs about anonymity, aggression, and the permanency of digital content. *Cyberpsychology, behavior and social networking*, 17.

Michelle Wright, Bridgette Harper, and Sebastian Wachs. 2018. The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition. *Personality and Individual Differences*, 140.

Jihad Zahir, Youssef Mehdi Oukaja, and Oumayma El Ansari. 2020. Arabic sexist comments detection in youtube: A context-aware opinion analysis approach. In *International Congress on Information and Communication Technology*, pages 461–469. Springer.