CoRoSeOf - An Annotated Corpus of Romanian Sexist and Offensive Tweets

Diana Constantina Höfels, Çağrı Çöltekin, Irina Diana Madroane
University of Tübingen, University of Tübingen, West University of Timișoara
diana-constantina.hoefels@student.uni-tuebingen.de, ccoltekin@sfs.uni-tuebingen.de, irina.madroane@e-uvt.ro

Abstract
This paper introduces CoRoSeOf, a large corpus of Romanian social media manually annotated for sexist and offensive language. We describe the annotation process of the corpus, provide initial analyses, and baseline classification results for sexism detection on this data set. The resulting corpus contains 39,245 tweets, annotated by multiple annotators (with an agreement rate of Fleiss' $\kappa = 0.45$), following the sexist label set of a recent study. The automatic sexism detection yields scores similar to some of the earlier studies (macro averaged F1 score of 83.07% on binary classification task). We release the corpus with a permissive license.

Keywords: corpus annotation, sexist language, offensive language, Romanian corpora

1. Introduction
Sexism is defined as a “prejudice or discrimination based on sex or gender, especially against women and girls”; and dates back from the “second-wave” feminism.\(^1\) The term was first used by Pauline Leet at an all-male lecture at a conservative college in 1965, where she talked about how women are positioned, marginalized, and considered invisible (Shapiro, 1985). Study after study shows how women are vulnerable to sexism: biases asserting that girls are more sensitive to pain and more reactive to it than boys (Cohen et al., 2014), a culture of sexism in the academia through the under-representation of women (Savigny, 2014), women at the workplace (Barnett, 2005; Verniers and Vala, 2018), and even in Natural Language Processing (NLP) models (Sun et al., 2019; Feine et al., 2020; Bhaskaran and Bhallamudi, 2019; Kiritchenko and Mohammad, 2018). Furthermore, a number of activists have used online platforms to raise awareness about and challenge gender injustice in the recent past, including Laura Bates (Bates, 2014), Reni Eddo-Lodge, Tarana Burke, and Lucy-Anne Holmes. Nearly 90% of the world’s population – both men and women – have prejudices against women (United Nations Development Programme, 2020). As a prevalent aspect of the world’s societies, it comes as no surprise that sexism permeates the language domain, with a debate among female activists dating back to the 19th century. Media studies, theory, and activism have challenged media representations of women, providing a platform for exploring social norms and expectations (Savigny, 2020). Sexist language is also used in social-media communication, leading to an environment of intimidation, insecurity, fear, and violence. Furthermore, sexist discourse on social media may facilitate the spreading, justification, and reinforcement of sexist acts. Thus, in order to protect the users of these platforms from sexist language, it is imperative to recognize, detect, and prevent sexist language in the online context. In this study, we contribute to this purpose by publishing a Romanian social media corpus manually annotated for sexism, and providing initial analyses of this corpus.

Romanian is a Balkan Romance language spoken by approximately 24 to 26 million people as a native language, while about 4 million speak it as a secondary language.\(^2\) Moreover, Romanian is spoken by immigrant communities in the United States, Canada, Australia, and many European Union countries. On the subject of Romanian corpora, we identified a lack of public corpora for classification tasks in an era when building linguistic resource collections is becoming increasingly in demand. Additionally, it seems that there are no linguistic corpora that pertain specifically to sexist language. Some of the most relevant Romanian corpora will be discussed in the following sections.

While offensive language does not have a universal established definition, it is typically associated with cursing, profanity, blasphemy, epithets, obscenity, and insults (Jay, 1992). Certain forms of offensive language, particularly hate speech and cyberbullying attracted considerable attention in recent years. Hate speech targeting gender, both women and LGBTQIA,\(^3\) has been the focus of some recent studies (Frenda et al., 2019; Pamungkas et al., 2020; Fersini et al., 2020). Sexist language is strongly related to (other) forms of offensive language use, and it has been one of the categories in a number of studies that take a broader view on hate speech or harassment directed to various target groups (Waseem and Hovy, 2016a; Kumar et al., 2021).

\(^1\)https://www.britannica.com/topic/sexism

\(^2\)https://www.britannica.com/topic/Romanian-language

\(^3\)Lesbian, Gay, Bisexual, Transgender, Queer and/or Questioning, Intersex, and Asexual and/or Ally
Within sexist language use, a distinction is often made between two forms sexism, hostile or benevolent (Glick and Fiske, 1996). Hostile sexism is typically explicit while benevolent is more subtle. For example, O femeie nu trebuie ațină nici măcar cu o floare, ‘A woman should not be touched, not even with a flower,’ is an expression used to emphasize the importance of women, highlighting the need to protect and value them. In spite of the positive outlook, it forms the perfect example of benevolent sexism, as it portrays women weaker than men and in need of male protection. As a result, the present study investigates both forms of sexist language, and their relation to the other forms or targets of offensive language.

A large and high-quality corpus is essential for training automated systems for detection and, hence, prevention of sexist language online. The main contribution of the present study is the introduction of CoRoSeOf, a large Romanian corpus annotated manually for sexist language on Twitter. The corpus is freely available with a permissive license at https://github.com/DianaHoefels/CoRoSeOf. We also provide initial computational experiments that may serve as a strong baseline for future studies that use the corpus for sexism detection. Since automatic methods tend to confuse different types of offensive language (Malmasi and Zampieri, 2018), the corpus is also annotated for non-sexist offensive language. Besides the use in automatic identification of sexist language use, such corpora can also be valuable for analyzing and understanding the sexist language use, for research in the humanities.

2. Related work

Sexism can manifest in a variety of forms, and researchers have been actively developing corpora and technologies to detect all its forms. According to Glick and Fiske (1996), coexistence of power differences between genders leads to hostility while strong interdependence leads to benevolent attitudes. In order to evaluate this theory, Jha and Mamidi (2017) designed a system to detect these two types of discrimination. Their data collection strategy consists of querying with phrases and hashtags commonly used to express benevolent sexism. Their own generated data set of tweets contains only benevolent sexism, therefore they employ an earlier data set (Waseem and Hovy, 2016b) to cover the hostile form of sexism. This data set contains 16k tweets classified into three classes: sexist, racist or neither. A shared task aimed at detecting hate speech targeting women, as well as immigrants, was SemEval-2019 – Multilingual Detection of Hate Speech Against Immigrants and Women (Basile et al., 2019b), where two widely spoken languages, English and Spanish, have been used to train and test participant systems. The tweets were collected through different approaches, such as filtering by keywords, monitoring hate accounts, and identifying and downloading haters’ posting history. The next study focuses also on hateful and sexist messages directed at women in both English and French (Chiril et al., 2019; Chiril et al., 2020). Data set consists of approximately 12k tweets. The novelty lies in the manner in which they investigate the sexist content, based on whether it is directed at women specifically, how women are described in general, or how women experience sexism in their everyday life. Other studies examine sexism from a multilingual perspective while still focusing on one form of sexism (e.g., misogyny (Fersini et al., 2020)).

Finer-grained approaches explore sexism outside of these two types. Rodríguez-Sánchez et al. (2020), with their first Spanish data set of manually annotated sexist tweets, the MeTwo data set. The data set is used in the first shared task for detecting sexism in social networks, EXIST (Rodríguez-Sánchez et al., 2021). Samory et al. (2021) focus also on many forms of sexist expressions by employing data from earlier studies (Waseem and Hovy, 2016b; Jha and Mamidi, 2017), as well as a new corpus of tweets they collected, ‘call me sexist’. Unlike some studies, which investigate sexism exhaustively (Rodríguez-Sánchez et al., 2021), others do so selectively, and examine just one form, an example being misogyny. Hewitt et al. (2016) tackled misogyny detection by collecting a rather small data set of tweets using keywords, which are analysed in a simple binary model. Despite its limitations, this study was still important in understanding in which ways sexist language is expressed on Twitter and the challenges associated with detecting it due to change in language usage. Another study describing detection of misogynistic language on Twitter is represented by Anzovino et al. (2018). The authors developed an annotation scheme for analyzing misogynistic content in detail. A keyword-driven approach is used by the data collection department to represent different misogyny categories and monitor the accounts of victims and misogynistic people.

Having emerged as a movement aimed at creating social change for victims of sexual harassment, the #MeToo movement has caused controversy, conversation and backlash from its inception. Modrek and Chakalov (2019) analyse English tweets with ‘#MeToo’ in the text, in order to quantify conversations surrounding the #MeToo movement, specifically first-person testimonies of sexual assault and abuse from both childhood and adulthood. Several other MeToo data sets were built in the past two years. The NLP community associates offensive language with a number of tasks. Typically, published studies
tackle specific tasks such as hate speech, racism, cyberbullying, discrimination, abusive language, while others approach it from a broad standpoint, and consider all its forms. Most of these studies, however, focus on English and this is partly due to the fact that most annotated data sets are in English. However, some recent studies have sought to detect hate speech in other languages, such as Dutch (Jha and Mamidi, 2017), French (Chiril et al., 2019; Papegnies et al., 2017), German (Ross et al., 2016; Strüß et al., 2019), Italian (Bosco et al., 2018; Poletto et al., 2017), Spanish (Anzovino et al., 2018; Basile et al., 2019b), or Turkish (Çöltekin, 2020). Other studies propose a multilingual approach to hate speech detection, e.g., English and French (Chiril et al., 2019), English and Spanish (Fersini et al., 2018) or even English, Italian, Polish, Portuguese and Spanish (Huang et al., 2020).

Some important shared tasks on hate detection include the Workshop on Abusive Languages Online (Waseem et al., 2017), the First Workshop on Trolling, Aggression and Cyberbullying (Kumar et al., 2018b), ICWSM 2020 Data Challenge on Online Safety (Founta et al., 2018), SemEval-2019 Task 5 (Basile et al., 2019b), GermEval Task 2, 2019–Shared Task on the Identification of Offensive Language (Strüß et al., 2019), and OffensEval 2020 (Zampieri et al., 2020).

There is a dearth of public Romanian corpora for many tasks and as previously mentioned an absence of corpora pertaining to sexist language. Some of the important corpora for Romanian include, the CoRoLa corpus (Mătălă et al., 2014), a reference corpus of contemporary Romanian, Evenimentul Zilei Corpus, RoCo-News (Tufiş and Irimia, 2006), the ROWAC Corpus and Romanian Word Skechtès (Macoveiciuc and Kilgarriff, 2010), MOROCO: The Moldavian and Romanian Di-lectal Corpus (Butnaru and Ionescu, 2019), and the first Romanian corpus of offensive language that has been recently released, ROFF - A Romanian Twitter Dataset of Offensive Language (Manolescu and Ionescu, 2019), and the recent Romanian Twit-2010), MOROCO: The Moldavian and Romanian D

3. Corpus collection and annotation

This section describes the methodology used for building CoRoSeOf, an annotated Corpus of Romanian Sexist and Offensive language, and the outcome of the annotation process.

3.1. Data Collection

Our data consists of tweets collected using Twitter streaming API4 between May 18th, 2021, and July 9th, 2021. A particular challenge in similar annotation tasks arises because of the fact that a random sample contains few examples of positive class (in our case, sexist tweets) in comparison to negative class. As a result a large number of tweets has to be annotated in order to obtain a reasonable number of offensive or sexist comments (Wiegand et al., 2019). Earlier studies have alleviated this problem by using strategies such as filtering by a set of representative keywords, names of people who are likely to be the victims or offenders, or even specific hash tags or events related to sexism (Chiril et al., 2020; Anzovino et al., 2018; Rodríguez-Sanchez et al., 2020; Fersini et al., 2018). However, this method of collection necessarily biases the data towards the keywords used.

Rather than committing fully to the task of minimizing bias, or maximizing the number of sexist instances, we collected tweets using two separate keyword lists. First, following Chiril et al. (2020), we used a set of keywords that are likely to occur in sexist or offensive language. Second, we used the most frequent Romanian words as the query terms. The first list increases the chance of positive instances, and negative instances that may be confusing due to specific keyword use. The second list gives a relatively unbiased, representative set of tweets.5

We collected Tweets that are identified as tweets in Romanian by the Twitter language detection system, and match both keyword lists (separately). The collection process yielded approximately 9.8 million matching the sexist keywords, and 207 million tweets matching the frequent Romanian words. From these two sets, we randomly sampled 40 000 tweets (75% based on sexist keywords, 25% based on general keywords).

Since we are interested in linguistic aspects of sexism, following earlier studies (Wiegand et al., 2019; Çöltekin, 2020), we filtered out the tweets matching the following criteria during sampling.

- retweets, including quote tweets
- tweets from verified users (these tweets tend to contain formal, verified content which are less likely to be sexist or offensive)
- tweets containing less than five alphabetic tokens (since they tend to include images, and little linguistic content)
- tweets that contain URLs (these are often advertisements, and the text without the associated links are not meaningful by itself)
- tweets that are not detected as Romanian (Twitter’s language detection mechanism produced frequent false positives)

3.2. Label set

We use a two-level labelling scheme. The first level is to divide content into two categories: ‘sexist’ or ‘non sexist’. In the second level, sub-types of sexist and non

---

4https://developer.twitter.com/en/docs

5The sexist keyword list was manually created by the first author. The frequency list is obtained from Romanian section of the Leipzig corpora (Goldhahn et al., 2012). Both lists are released at the corpus web page at https://github.com/DianaHoefels/CoRoSeOf.
sexist content is annotated. For sexist content, follow-
ing Chiril et al. (2020), we use three categories: ‘sexist
direct’, ‘sexist descriptive’, and ‘sexist reporting’. For
non-sexist posts, we also distinguish ‘non-offensive’,
and ‘offensive’ (but not-sexist) posts. With the addi-
tion of a special label, which indicates that the annot-
ator could not decide, we label each tweet with one
of the following mutually-exclusive labels. We list the
label set with brief descriptions, and examples below.
The annotation guidelines provided to the annotators is
published at the corpus repository. ¡

Sexist direct The tweet includes sexist ele-
ments and is addressed directly to a particular gen-
der, typically women or groups of women.
(1) a. Chiar crezi că arătă bine? Mai uită-te
odată în oglindă, si răspunde din nou.
Chiar crezi că ești suficient de slabă?
#roedwt.
‘Do you really think you look good? Take
another look in the mirror, and answer again.
Do you really think you’re thin enough?
#roedwt.’ 6
b. Tu cât de proastă ești, nu înțelegi nici
tenisul, nici fotbalul. Ești bună la vopsit
părul.
‘How stupid you are, you don’t understand
tennis or football. You’re good at dyeing your
hair.’

Sexist descriptive The tweet describes one or
more people, typically a woman or women, in a
sexist manner, without addressing them directly.
(2) a. Alexandra Stan e genul de femeie care
dacă ar fi fost urâtă și fără succes o
plăceau toate celelalte femei.
‘Alexandra Stan’ is the kind of woman who,
if she had been ugly and unsuccessful, would
have been liked by all the other women!’

b. Despre Viorica Dăncilă: “Dacă apare o
noață banchetă de 100 lei, va scrie pe ea
una tută lei.”
About Viorica Dăncilă: 8 “If a new 100 lei ban-
knote is issued, one stupid lei will be printed
on it.”

Sexist reporting The tweet reporting an act of
sexism witnessed or heard from other sources.
(3) a. E triggering. Mai ales când primesc o
cerere în toată noptii de la un bărbat.
‘It’s triggering. Especially when I get an invi-
tation from a man in the middle of the night.’
b. ...Azi au dat alt câz de o femeie ucisă,
au ajuns la 20 de la începutul anului.
Mare parte din vină o are justiția.
‘Today they reported another case of a mur-
dered woman, they’ve reached 20 since the
beginning of the year. Justice is largely to
blame.’

Non-sexist offensive The tweet has no sexist
connotations, but contains offensive language
(4) Ce bine vă stă împreună, doi prosti amândoi.
‘You two look good together, two fools.’

Non-sexist (and non-offensive) there are no sexist
or offensive elements in the tweet, nor does the
tweet carry any sexist or offensive connotations.
Tweets may contain sexist or offensive elements
or hashtags, (providing little or no context) but the
overall meaning of the tweet is neither offensive
nor sexist.
(5) Nu-i așa că e superbă?.
‘Isn’t she gorgeous?’

Cannot decide The tweet lacks context, contains
sexist undertones, but cannot be unambiguously
tagged with one of the labels above.

3.3. Annotation Process
Annotation was carried out by a team of ten annotators
(seven female and three male), with reasonably similar
background: all native speakers of Romanian, university
students majoring in Languages and Literature and
Modern Applied Languages, ranging in age from 20 to
25 years old, and with an interest/knowledge in gender
studies. Their contribution was rewarded with univer-
sity credit points. Before the actual annotation task, the
annotators went through a training period where they
attended an orientation session, followed by a quiz to
confirm their understanding of the concepts, annotation
of 50 sample tweets and discussion of disagreements.
The annotation task was conducted remotely over the
period of a month, during which we held regular on-
line meetings where problematic tweets were reviewed
and discussed, making sure that the labels reflected the
annotation scheme.
In the annotation phase, we used Google forms for an-
notating 40K tweets. We split the data randomly into
160 forms each containing 250 tweets, and assigned
each form to three annotators. Since the annotators
committed to working for a different number of hours,
the number of forms assigned to annotators differ.
At the end of the annotation process, all tweets received
three annotations. We automatically determined the
consensus label using majority vote. In case of three
different labels, the final label was assigned by the first
author of the paper. A total of 1457 cases of non-
agreement were registered, and the final labels were
adjudicated by selecting a label among those assigned
by the annotators, or by selecting an entirely different

6 A community on Twitter where everyone has an eating
disorder.
7 Romanian singer-songwriter, known internationally for
her single ‘Mr. Saxobeat’.
8 Former Prime Minister of Romania. During her mandate
as prime minister of Romania, she was mocked for a series
of mistakes that some said were proofs of her poor language
skills (both English and Romanian).
one based on the annotation guidelines. The data set includes both the original version (No agreement and Cannot decide labels included), and the updated version (final labels are provided to the non-agreement and ambiguous samples).

After removing duplicates and tweets that do not contain linguistic content or those that are not in Romanian, the final data set contains 39,245 tweets.

### 3.4. Annotation results

#### Overall statistics

The label distribution in the final corpus is presented in Table 1. We observe that tweets that are offensive and sexist are distributed similarly, 9.93% are sexist, 11.01% offensive. While keyword-based sampling allowed us to use the valuable annotator time more efficiently, the random sampling provides a relatively unbiased picture of sexist and offensive language use in Romanian Twitter.

Not surprisingly, the proportion of sexist tweets is higher for the tweets that are sampled using sexist keywords 9.41% in comparison to 6.48% in the random tweet collection. Although the difference between the percentages may seem small, we would need to annotate approximately 50% more data using random sampling to get the same amount of sexist tweets. We also observe a similar, but less pronounced trend in the ratio of offensive tweets in two different samples (10.70% in sexist keyword-based sampling in comparison to 8.01% in random set).

In comparison to the most similar study, the French sexist language corpus by Chiril et al. (2020), our corpus is considerably larger. The French corpus contains 12,274 tweets. However, the French corpus contains a higher number of positive labels, 4,487 sexist samples compared to our 3,897 positive labels, in our corpus. Comparing the distribution of labels for sexism in the two corpora, CoRoSeOf has a higher level of sexist direct while the French corpus has a higher rate of sexist reporting and the lowest amount of sexist direct while the French corpus has a higher level of sexist and offensive language use in Romanian Twitter.

Most disagreements that are resolved by majority vote are between Non sexist (and non offensive) and Non sexist offensive, constituting 48.29% of all disagreements (e.g., a. Occidentul a comis genocid pe oriunde s-a dus să colonizeze!! ‘The West committed genocide wherever it went to colonize!!’ b. Toca se poartă de către studenti, nu de cei care se tem de BAC! ‘The cap is worn by students, not by those who fear the BAC!’). Disagreements between Non sexist and Sexist direct constitute 9.94% (e.g., Bună seara frumoasă!, ‘Good evening beautiful!’) and disagreements between Non sexist and Sexist descriptive constitute 8.92% of the disagreements (e.g., Toate fetele fac paste si zic că stiu să gătească. ‘All girls cook pasta and say they know how to cook.’).

The samples that were not resolved by majority vote were determined by the first author of this paper. For ambiguous samples, the percentage of corrections which resulted in sexist labels is 34.69%, 57.14% for non-sexist labels and a small number of samples deemed offensive. Disambiguation difficulties of the sexist tweets may have been caused by incorrect interpretations, overlooking some elements of a tweet, such as emojis, usernames, punctuation, or perhaps lack of language use and world knowledge. In addition, some of the reasons listed above could also apply to the samples that, after correction, were deemed non-sexist. Further, there could have been confusion due to the sexist terms found in the non-sexist samples, although these terms were neutrally employed. With re-

| Label          | Sampling |
|----------------|----------|
|                 | Keyword | Random | Total  |
| Sexist          | 3 332   | 565    | 3 897  |
| Direct          | 1 838   | 333    | 2 171  |
| Descriptive     | 1 301   | 206    | 1 507  |
| Reporting       | 193     | 26     | 219    |
| Non-sexist      | 28 183  | 7 165  | 35 348 |
| Offensive       | 3 639   | 681    | 4 320  |
| Not offensive   | 24 544  | 6 484  | 31 028 |

Table 1: Label breakdown in CoRoSeOf including the number of labels per sampling source.

*κ* scores, Rodríguez-Sanchez et al. (2020) report an average of 0.75 (Cohen’s *κ* for binary annotation, sexist vs. non-sexist).

Looking further into detailed annotator agreement, the annotators unanimously agree on a label for 27,060 (68.89%) of tweets. In 10,751 (27.37%) of the cases, two annotators agree on the label, leading to majority decision. All three annotators disagree on 1,457 (3.71%) of the tweets.

The final data set contains 1,301,206 tweets. The samples that were not resolved by majority vote were determined by the first author of this paper. For ambiguous samples, the percentage of corrections which resulted in sexist labels is 34.69%, 57.14% for non-sexist labels and a small number of samples deemed offensive. Disambiguation difficulties of the sexist tweets may have been caused by incorrect interpretations, overlooking some elements of a tweet, such as emojis, usernames, punctuation, or perhaps lack of language use and world knowledge. In addition, some of the reasons listed above could also apply to the samples that, after correction, were deemed non-sexist. Further, there could have been confusion due to the sexist terms found in the non-sexist samples, although these terms were neutrally employed. With re-

A Bacalaureat exam (or BAC as it is known in Romania) is taken by high school students in order to graduate.
spect to the samples in which each annotator assigned a different label, the correction results reveal that sexist labels are predominant with 37.54% (Sexist descriptive samples are the most prevalent), followed by Non-sexist offensive with 32.81%, and Non sexist (non offensive) with 29.65%. The general observation is that in the first round of annotation, the majority of tweets are labelled as either sexist or offensive. The second round of annotations revealed the greatest amount of indecision, while the last round indicated a preference for non-sexist labels. However, in terms of sexist labels, an almost equal number of tweets are labelled as sexist in all rounds of annotations; however, the annotators hold varying views on what is sexist and what is not. The percentage of corrections in which we agree with one of the three annotators, for non agreement samples was 88.13% of the time; in the other cases, we assigned a different label that does not correspond to any of the existing three. In the following example, Duaaa... Sper cà sunt doar ocupat si nu au probleme ‘Yeah... I hope they’re just busy and not in trouble.’ annotators viewed the tweet as either Cannot decide, Sexist descriptive, or Sexist direct. The first annotator’s indecision is understandable given the tweet is short and devoid of any context. In contrast, we could not identify any sexist terms to justify the annotations of the next two annotators, so we marked this tweet as Non sexist. Some of the other examples in which we disagree are tweets that contain derogatory terms, but are deemed sexist instead of offensive. Further, a number of tweets reporting sexist acts were labelled as either Sexist direct or Sexist descriptive. This could indicate that annotators were either uncertain, or inattentive, in making distinctions between the sexist forms.

Differences based on annotator gender  An interesting aspect of our corpus is the fact that we had both male and female annotators. During the annotation process, we received 76,324 annotations from the female annotators and 41,513 from the male annotators. The distributions of labels within groups are generally similar. However, in general, the male participants provided a slightly higher rate of sexist annotations. On the other hand, female annotators found a proportionally more non-sexist tweets offensive (13.28% vs. 9.61%), and they also showed a slightly higher rate of indecision (label Cannot decide) in comparison to male annotators (1.40% vs. 0.92%). The distribution of labels based on annotator gender is presented in Table 2.

|                | Female |       | Male |       |
|----------------|--------|-------|------|-------|
|                | n      | %     | n    | %     |
| Sexist         |        |       |      |       |
| Direct         | 4147   | 5.43  | 2756 | 6.64  |
| Descriptive    | 2916   | 3.82  | 1894 | 4.56  |
| Reporting      | 416    | 0.55  | 391  | 0.94  |
| Non-sexist     |        |       |      |       |
| Offensive      | 10136  | 13.28 | 3989 | 9.61  |
| Not offensive  | 57641  | 75.52 | 32102| 77.33 |
| Cannot decide  | 1068   | 1.40  | 381  | 0.96  |

Table 2: Label distribution based on annotator gender.

4. Baseline Classification Experiments
An important use case for the corpus presented here is automatic identification of sexist language. To provide a baseline for future studies using the corpus for detecting sexist tweets, we run a set of experiments with straightforward text classification methods. Since inconsistent annotations are one of the sources of failures for the machine learning models, these experiments also provide evidence for the consistency of the annotations.

We use linear SVMs with sparse character and word n-gram features, to obtain our baseline results. We use both character and word n-grams, both concatenated as a single feature matrix, and weight them using tf-idf. We experiment with both binary classification (sexist–non-sexist), and 3-way classification to sexist sub-categories (direct, descriptive, reporting).

For all results reported below, we consider the maximum character (in range [0, 7]) and word (in range [0, 4]) n-grams, the SVM margin/regularization parameter ‘C’ (in range [0.001, 5.0]), and a preprocessing parameter specifying whether to perform case normalization on word n-grams, character ngrams, both, or none. We use class weights inversely proportional to class frequencies to counteract label imbalance. We run 10-fold cross validation experiments on the whole data for 3,000 random samples from the above hyperparameter space, and report the average scores over 10 folds for the hyperparameter setting with the highest macro averaged F1 score.

4.1. Classification Results
The best scores we have obtained for binary classification (sexist–non-sexist) is 83.14% (sd=1.01%) F1 score, 83.07% (sd=0.75%) precision and 83.24% (sd=1.48%) recall. These scores are similar (but substantially better than) the best scores (76.2% macro averaged F1 score with pre-trained language models) presented by Chiril et al. (2020) on the French sexist language corpus. Part of the performance difference may be explained by the fact that our corpus is substantially larger. Another contributing factor can be the sampling method. Since the French corpus is collected mainly based on sexist keywords, it is likely that the negative examples also include sexist keywords, making the task more challenging. Nevertheless, this clearly shows that the CoRoSeOf annotation is consistent enough for a machine learning method to be able to find necessary features also.
signal for detecting sexist language. As expected, the rate of false positives is higher for offensive tweets in comparison to non-offensive ones. The classifier identifies 4.69% of the offensive tweets as sexist, while the percentage of non-offensive tweets mistakenly identified as sexist is 2.81%. A manual inspection of the predicted samples, however, revealed that a considerable amount of false positives and false negatives turned out to be true positives, indicating that there were some inconsistencies in the annotation process. In terms of false negatives, the model incorrectly predicts the negative class due to the presence of prominent women and men usually targeted for sexism, such as politicians, athletes, and artists, plus the absence of explicit sexist keywords. (e.g., [...] Utitâ-vâ pe activitatea ei de senatorat, este aproape zero în materie de inițiatii legislative, face circ pentru rating si atât. [...] ‘Look at her work as a senator, it is almost zero in terms of legislative initiatives, she makes a circus for ratings and that is it.’ Comparing the annotator agreement with the classifier decision, we observe that the classifier makes fewer mistakes classifying the instances that were agreed by the annotators unanimously (2.34%), while classifier decision degrades substantially for the instances whose labels were determined by majority, resulting in errors in 11.24% of the cases. Finally, although lower than the binary classification, as expected, the three-way classification into sexist language types also results into rather good scores. The classifier obtains macro-averaged precision, recall and F1 scores of 71.62%, 69.29%, and 70.02%, respectively.

5. General discussion

The previous sections introduced CoRoSeOf, a social media corpus of Romanian annotated primarily for sexist language. To the best of our knowledge, this is the first annotated collection of Romanian sexist language. A recent paper introduced a corpus of Romanian offensive language (Manolescu and Ciîltekin, 2021), providing detailed categories of offensive language without the explicit marking of sexism. Our offensive (non-sexist) language annotation may be complementary to the usage of this corpus. Since we follow the annotation schemes of similar recent studies, the present corpus is also a valuable resource for multi-lingual and cross-lingual analysis or automatic detection of offensive and sexist language.

Each text in CoRoSeOf is annotated by multiple annotators, yielding reasonable agreement scores considering the difficulty and subjectivity of the task and the agreement scores obtained in earlier studies. To our knowledge, CoRoSeOf is also the largest corpus of social media with sexist language annotations for any language we are aware of. Furthermore, our hybrid sampling technique combining random sampling and sexist keywords allows studying a more representative sample of the source data, while increasing the number of positive examples in the resulting corpus.

5.1. Qualitative analysis

To discover in what way gender discrimination occurs and which group is most affected by it, we analyse the sexist and offensive tweets from CoRoSeOf. The label distribution in the corpus shows that sexist language is prevalent on Romanian Twitter; furthermore, the highest percentage is represented by sexist direct, compared with a much lower number of instances describing women in a sexist manner, and, lastly, the reported sexist acts are surprisingly rare. We manually inspected the data set so as to comprehend the various manners sexism is articulated on Romanian Twitter. From a review of samples labelled as sexist direct, the most representative type of sexism within our corpus, we identify that it is expressed in both a benevolent and hostile manner. We find tweets that sound complimentary, such as women are beautiful, delicate flowers, or princesses who need protection from men, as well as tweets that sound hostile and derogatory, such as women are stupid or incapable. Furthermore, objectification, stereotyping, and hostility toward women can occur all at once, for example:

(6) Du-te la bucătărie curvă travestită ce esti. ‘Go to the kitchen you transvestite bitch.’

Within each sexist label subset, we identified the most frequent sexist terms in CoRoSeOf or the terms that contributed most to generating sexist content. Appendix A provides a list of the most frequently used sexist terms in Romanian according to each category examined in this paper. Sexist direct language is characterized by words that refer to women in a sexually objectifying manner. While the highest ranking word for this category, the adjective frumoasă, ‘beautiful’, sounds positive, it is often used to discriminate against women based on their appearance. Moreover, the adjective bună, whose literal meaning is ‘good’, is mostly used to describe the sexual characteristics of a person, most commonly women. Similarly, the noun suflet, ‘soul’, is used to describe a person’s affective, intellectual and volitional processes. However, we found examples in our data set where it is used to describe a body part as in (7).

(7) Ce suflet frumos ai! ‘What a beautiful soul you are!’

‘What beautiful breasts you have!’

In the example above, we provide the literal translation, followed by the one expressing the intended meaning. Overall, the sexist tweets sub-categorized as direct, represent women as sexual objects, with a hand-ful being aggressive toward women. The same process of sexually objectifying women is at work through the Sextist descriptive tweets. This collection is extremely frequent with the noun, femeie, ‘woman’. Women are described as both sexual objects and intellectually challenged, as well as praised or despised for their physical
appearance. In the last category of sexism (8) the sexual objectification of women is reported through stories and mentions of cases of both physical and verbal abuse, for example:

(8) a. Saptămâna asta eu cu câteva colegie am iesit afară în parcul să ne luăm mâncare ... si cum mergeam noi pe stradă am auzit unu în spatele nostru zicând, 'mamă ce bulane'.

'This week a few colleagues and I went out to grab lunch, and as we walked down the street we heard someone behind us say, 'Dang, you’ve got beautiful legs.''

b. Este acum la televizor o femeie care a venit pentru că a bătut-o sotul [...]'

'There is a woman on TV now who came in because her husband beat her [...]'

These are reports of classic examples of hostile sexism, (8a) uses of sexist language or insults, while (8b) reports a case of domestic violence.

Contrary to the frequency of sexist tweets directed at women, there are a handful of instances where we observed sexism directed at men. In such instances, women mirror the same behaviour as men, namely, they are aggressive, make inappropriate comments to sexually objectify them, or present them as liars, insensitive, frivolous, etc. For example:

(9) Telecomanda si bărbatul se repară cu pumnul.

'The remote control and men are to be repaired with the fist.'

Upon closer inspection, we find a few samples in our collection that contain fabricated forms of sexism, such as nonsensical combinations of sexual slurs, generated by an algorithm. While we do not provide examples, the reader can infer that they contain extreme explicit content such as sexual or pornographic.

Some of our annotators suspected these types of tweets to be generated by Twitter bots. The purpose of Twitter bots, also known as zombies, is to create a massive stream of tweets in order to achieve specific goals on a large scale. In their quest to accomplish certain goals, some of these bots are sexist and can be used for political purposes or to influence elections by posting sexist content about female candidates, as was observed during Hillary Clinton’s 2016 election campaign.11 However, we remain uncertain about the source of the fake sexist tweets, since we could not reliably identify them and our conclusions simply come from the frequency of accounts that we found.

After reviewing the offensive tweets, we have discovered that they address mostly topics such as politics, cryptocurrency, gaming, and Covid-19. There is a wide variety of insults in these tweets, targeting single users, politicians, politicians’ parties, businessmen, or entire social groups. Moreover, we also detected foul language expressing antisemitism, profanity, racism, and homophobia. Appendix B lists the most common terms used to express offensive language in Romanian.

(10) Åstia sunt retardatii votati de niste trogloditi de dreapta.

'These are the retards voted in by some right-wing troglodytes.'

A number of cases we observed exhibited irony and sarcasm. The example below illustrates sarcasm, in which a user calls a person garbage.

(11) Dar chiar si în Germania am observat un gunoi. Îl vezi?

'But even in Germany I noticed garbage. Do you see it?'

Our evidence indicates that sexism most often shows up in an overt, blatant manner, expressed in both caring and nurturing, but also harassing and threatening ways, and it is mostly directed at women. Mills (2008) points out that sexism has become mostly indirect because of the success of the feminist language reform and political correctness policies. The rise of the online environment seems to have led to a reversal. A similar observation can be made regarding the use of offensive language, namely the words are used openly and forthrightly. It appears, however, that irony and sarcasm occur more frequently in offensive than in sexist language. While sexism is predominantly addressed to women, offensive language targets a broader range of groups and individuals.

6. Conclusions and Future Work

In this paper, we introduce a new dataset that can be used for analysis and automatic detection of sexist and offensive language. Our best knowledge indicates that this is the first annotated corpus of offensive words in Romanian, and the largest corpus ever annotated for sexist terms in any language. Each tweet in our corpus is annotated by three annotators, and the annotator agreement and initial classification experiments indicate the consistency of annotations. We believe the present corpus may aid researchers interested in studying sexist language, and the development of NLP systems for detecting sexist language in online communication. There are a number of different tests and experiments still to be carried out in the future. A revised and improved version of the corpus will be developed to capitalise on the takeaways of the current methods and propose alternative approaches. It is intended that the future research will explore both overt and subtle forms of sexism, and that a quantitative analysis of the new dataset will be conducted in greater detail. Furthermore, we aim to collect data with fewer filters (e.g., verified users, since some of them are focused on spreading sexist and offensive language).

11https://www.theatlantic.com/technology/archive/2016/11/election-bots/506072/
7. Acknowledgements

We would like to thank the team of annotators from the Interdisciplinary Center for Gender Studies - West University of Timisoara, for their extensive work and dedication on this project (in alphabetical order), Anamaria Andrei, Raluca Ardeanu, Edward Bojboi, Octavia Cofocaru, Cristina Giurcă, Costel Olaru, Roberta Recalo, Diana Stanciu, Tiberiu Tomescu and Carmen Tuns.

8. Bibliographical References

Anzovino, M., Fersini, E., and Rosso, P. (2018). Automatic identification and classification of misogynistic language on Twitter. In Max Silberztein, et al., editors, Natural Language Processing and Information Systems, pages 57–64, Cham. Springer International Publishing.

Barnett, R. (2005). Ageism and sexism in the workplace. Generations (San Francisco, Calif.), 29:25–30, 09.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019a). 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, June. Association for Computational Linguistics.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019b). 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, June. Association for Computational Linguistics.

Bates, L. (2014). Everyday Sexism. Everyday Sexism. Simon & Schuster.

Bhaskaran, J. and Bhallamudi, I. (2019). Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 62–68, Florence, Italy, August. Association for Computational Linguistics.

Bosco, C., Dell’Orletta, F., Poletto, F., Sanguinetti, M., and Tesconi, M. (2018). Overview of the EVALITA 2018 hate speech detection task. In EVALITA 2018—Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, pages 67–74, 01.

Butnaru, A. M. and Ionescu, R. T. (2019). MOROCO: The Moldavian and Romanian Dialectal Corpus. ArXiv.

Çöltekin, c. (2020). A corpus of turkish offensive language on social media. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 6174–6184, Marseille, France.

Chiril, P., Benamara, F., Moriceau, V., Coulomb-Gully, M., and Kumar, A. (2019). Multilingual and Multitarget Hate Speech Detection in Tweets. In Conférence sur le Traitement Automatique des Langues Naturelles (TALN - PFIA 2019), pages 351–360, Toulouse, France, July. ATALA.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., and Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in French tweets. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1397–1403, Marseille, France, May. European Language Resources Association.

Chukwuere, J., Chukwuere, P., and Khumalo, N. (2018). Cyber bullying of female students: An exploration of literature study. Gender and Behaviour, 15:9983–9995, 03.

Cohen, L. L., Cobb, J., and Martin, S. R. (2014). Gender biases in adult ratings of pediatric pain. Children’s Health Care, 43(2):87–95.

Feine, J., Gnewuch, U., Morana, S., and Maedche, A. (2020). Gender bias in chatbot design. In International Workshop on Chatbot Research and Design, pages 79–93, 01.

Fersini, E., Rosso, P., and Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. In Ibereval@SEPLN.

Fersini, E., Nozza, D., and Rosso, P. (2020). AMI @ EVALITA2020: automatic misogyny identification. In Valerio Basile, et al., editors, Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of CEUR Workshop Proceedings. CEUR-WS.org.

Fleiss, J., Cohen, J., and Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 72:323–327.

Founta, A.-M., Djouvas, C., Chatzakou, D., Leondiatis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In 11th International Conference on Web and Social Media, ICWSM 2018. AAAI Press.

Fredda, S., Ghanem, B., Montes-y Gómez, M., and Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. Journal of Intelligent & Fuzzy Systems, 36(5):4743–4752.

Glick, P. and Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. Journal of Personality and Social Psychology, 70:491–512, 03.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pages 759–765, Istanbul, Turkey, May. European Language Resources Association (ELRA).
Manolescu, M. and Çöltekin, C. (2021). Recent advances in natural language processing (ranl), 2021. In Proceedings of Recent Advances in Natural Language Processing. Online event, September 7th, 2021, page 899–904.

Mills, S. (2008). Language and Sexism. Cambridge University Press.

Mititelu, V. B., Irimia, E., and Tufis, D. (2014). CoRoLa — the reference corpus of contemporary Romanian language. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 1235–1239, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Modrek, S. and Chakalov, B. (2019). The #metoo movement in the united states: Text analysis of early twitter conversations. Journal of Medical Internet Research, 21, 09.

Muttaqin, M. and Ambarwati, N. (2020). Cyberbullying and woman oppression. In 6th International Conference on Social and Political Sciences (ICOS-APS 2020), pages 545–553, 01.

Pamungkas, E. W., Basile, V., and Patti, V. (2020). Misogyny detection in twitter: a multilingual and cross-domain study. Information Processing & Management, 57(6):102360.

Papegnies, E., Labatut, V., Dufour, R., and Linarès, G. (2017). Detection of abusive messages in an online community. In 14ème Conférence en Recherche d’Information et Applications (CORIA), Conférence en Recherche d’Information et Applications, pages 153–168, Marseille, France, March.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., and Bosco, C. (2017). Hate speech annotation: Analysis of an italien twitter corpus. In Clic-it.

Rodríguez-Sánchez, F. J., Carrillo-de Albornoz, J., and Plaza, L. (2020). Automatic classification of sexism in social networks: An empirical study on twitter data. IEEE Access, 8:219563–219576.

Rodríguez-Sánchez, F., de Albornoz, J. C., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., and Donoso, T. (2021). Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural, 67(0).

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Michael Beißwenger, et al., editors, Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, pages 6–9.

Samory, M., Sen, I., Kohne, J., Flocck, F., and Wagner, C. (2021). "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples.

Savigny, H. (2014). Women, know your limits: cultural sexism in academia. Gender and Education, 26(7):794–809.

Savigny, H. (2020). Sexism and misogyny. In The International Encyclopedia of Gender, Media, and Communication, pages 1–7. Wiley Online Library, 07.

Selkie, E. M., Kota, R., and Moreno, M. A. (2016). Cyberbullying behaviors among female college students: Witnessing, perpetration, and victimization. College student journal, 50 2:278–287.

Shapiro, F. R. (1985). Historical notes on the vocab-
ulary of the women’s movement. *American Speech*, 60:3.

Struš, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., and Klenner, M. (2019). Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS) 2019*, pages 352 – 363, München [u.a]. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review.

Tufiş, D. and Irimia, E. (2006). RoCo-News: A hand validated journalistic corpus of Romanian. In *Proceedings of the 5th LREC Conference*, pages 869–872, 01.

United Nations Development Programme. (2020). Tackling social norms: A game changer for gender inequalities.

Verniers, C. and Vala, J. (2018). Justifying gender discrimination in the workplace: The mediating role of motherhood myths. *PLOS ONE*, 13:e0190657, 01.

Waseem, Z. and Hovy, D. (2016a). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (2016b). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Zeerak Waseem, et al., editors. (2017). *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop at KONVENS 2018*, pages 1–10.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Piteois, Z., and Çöltekin, c. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.
### A. Sexist words frequency

Listed here are the Romanian terms and their English translation, that carry the most sexist content or help build it. As with all Romance languages, Romanian is an inflected language, therefore most of the verbs in this list are changed to their infinitive forms, though some are left verbatim, since the tense they were found in carries meaning to the type of sexism. These are only rough estimates of the sexist language's most prominent words within the corpus (some inflected words may have been missed).

| Direct            | Reporting         | Descriptive       |
|-------------------|-------------------|-------------------|
| frumoasă beautiful| femeie woman      | femeie woman      |
| bună hot           | bărbat man        | zice say          |
| pulă dick          | fată girl         | bărbat man        |
| vrea want 292      | pula dick         | ea she            |
| cur sexy 236       | spune say         | fată girl         |
| sexy sexy 235      | frumoasă beautiful| eu l              |
| pizdă twat 220     | ea she            | băntuți beaten    |
| dulce sweet 191    | fute screw        | despre about 13   |
| plicăca like 166   | cur butocks       | el he             |
| esti you are 149   | sex sex           | crede believe 12  |
| linge lock 140     | arăta to look     | văzut seen 12     |
| pup kiss 140       | pizdă twat        | agresată abused   |
| fute screw 122     | curvă whore       | permite allow 11  |
| superbă beautiful | femeie woman      | vrea want 11      |
| sex sex 114        | ele they          | fute screw 9      |
| sâni breasts 138   | bani money        | fapt fact 8       |
| arăta to look 91   | iubită girlfriend | dreptate justice 8|
| gară mouth 89      | suge suck         | merită deserve 8  |
| femeie woman 75    | proastă stupid    | scurtă short 8    |
| fund butocks 72    | nevastă wife      | sot husband 8     |
| păsărică pussy 72  | iubi love         | vină fault 8      |
| suge suck 72       | sexy sexy         | arăsă burnt 8     |
| limbă tongue 68    | despre about 21   | vina ta your fault|
| bagă stick 64      | picoare legs      | fustă skirt 8     |
| drăgătă pretty 55  | urătă ugly        | jigni offend 7    |
| succulentă succulent 54 | doamnă Ms 20 | sexism sexist 7   |
| doamnă Ms 45       | simte feel        | iubi love 7       |
| perfectă perfect 43| târă bitch        | sex sex 7         |
| fată girl 42       | sexual sexual     | glumă joke 6      |
| iubi love 40       | drept right       | frumoasă beautiful|
| spermă sperm 36    | pat bed           | copii children 6  |
| iubire love 33     | scurtă short      | coleg coleg 6     |
| dragă bear 33      | gura mouth        | corp body 6       |
| dulceată sweet 33  | mamă mother       | accepta agree 5   |
| noapte night 33    | feminină female   | mamă mother 5     |
| corp body 32       | coaie balls       | moartă dying 5    |
| iubită girlfriend 30| fund butocks      | scuze excuse 5    |
| wow wow 30         | fustă skirt       | victimiza victimize5|
| pozitie position 29| muiere woman      | viol rape 5       |
| mamă mother 28     | gelouă jealous    | bani money 4      |
| buze lips 27       | grasă fat         | acuzatie accusations|
| foc fire 27        | machiaj makeup    | mentalitate mentalită 4|
| bărbat man 24      | pulnă fat         | considerată considered 4|
| picioare legs 24   | amântat mistress 8| urată ugly 3     |
| nebunăcăcă crazy 22| suparată upset    |                 |
| apetisantă appetizing 21 | tate tits 8 |                 |
| excita excite 19   | corp body 8       |                 |
| curvă whore 18     | arogantă arrogant |                 |
| floare flower 17   | misogin misogynist|                 |
| printese princess 16| compliment       |                 |
| joc game 15        | porno porn        |                 |
| dragoste love 14   | blondă blonde     |                 |
| păroșta stupid 14  | salariu salary    |                 |
| sfârșit nail 14     | drăgătă pretty    |                 |
| freca rab 12       | creier brain      |                 |
| jucă play 12       | decolteu decolletage |                 |
| sclav slave 12     | bătaie beat       |                 |
| buci butocks 11    | spermă sperm      |                 |
| fierbinte hot 11    | speria scare      |                 |
| goală naked 11     | inteligentă intelligent|            |
| porno porn 11      | feministă feminine|                 |
| provocatore provacative 11 | prostituată prostitute | |
B. Offensive words frequency

Listed here are the Romanian terms and their English translation, that carry offensive content or help build it. These are only rough estimates of the offensive language's most prominent words within the corpus (some inflected words may have been missed).

| Term       | English     | Frequency |
|------------|-------------|-----------|
| pula       | dick        | 1282      |
| coaie      | balls       | 238       |
| bag        | stick       | 193       |
| cur        | buttocks    | 192       |
| prost      | stupid      | 132       |
| drac       | devil       | 125       |
| plm        | cuss        | 91        |
| taci       | shut up     | 64        |
| morti      | dead people | 50        |
| muie       | fellatio    | 47        |
| sex        | sex         | 46        |
| fute       | screw       | 39        |
| pizda      | twat        | 40        |
| Dumnezeu   | God         | 33        |
| cringe     | cringe      | 30        |
| gay        | gay         | 27        |
| mamă       | mother      | 27        |
| evrei      | Jewish      | 42        |
| Lucifer    | Lucifer     | 23        |
| ticălos    | skunk       | 21        |
| mizerie    | filth       | 18        |
| tradator   | traitor     | 16        |
| mincinos   | liar        | 16        |
| covid      | covid-19    | 13        |