Corpus of Romanian Sexist and Offensive Language

Annotation guideliness

Author: Diana Höfels
diana-constantina.hoefels@student-uni.tuebingen.de

Academic supervisor: Dr. Çagri Cöltekin
ccoltekin@uni-tuebingen.de

Eberhard Karls University of Tübingen — July 4, 2021

Introduction

This document represents a guide that presents aspects of the sexism annotation task with the purpose of creating a **corpus of sexist and offensive language in Romanian.**

Content disclaimer Please pay special attention to the following information. This guide contains examples that include offensive language, forms of expression that incites, promotes or justifies hatred, violence and discrimination against a person or group of people for various reasons.
Similarly, we warn you that the text data prepared for the annotation task, namely tweets from the social networking platform Twitter, may contain violent, offensive, brutal language, insults, threats, words that incite violence, degrading, humiliating or discriminatory expressions. Please be mindful of the fact that the sole goal of this task is to build a corpus, and do not use the information in the data collected to incite or spread hatred of any type and form towards Twitter users, existing in this data set.

Social media users are exposed to various types of abusive behavior, such as hate speech, cyberbullying, racist and sexist comments. Offensive language is comprised of swearing, cursing, dirty, obscene, insulting words, or slander. Some of these offensive words are directed at a single gender and are mainly used to signal behaviors that are not in line with the ideas of masculinity or femininity of a society (Kremin, 2017). Sexism is defined as a **prejudice or discrimination based on sex, especially discrimination against women** [1] and it is found in all areas of life. Acts of sexism may seem harmless, however it can create a climate of intimidation, fear and insecurity, which leads to the acceptance of violence, especially against women and girls. Acts of sexism are also found in the online environment, where discussions are not always objective or constructive, and can easily escalate into hatred and discrimination (Chiril et al., 2020).

In line with our knowledge, a corpus of Romanian sexist phrases does not currently exist, therefore we attempt to create a corpus that aims at quantitative scientific research and could potentially be used in the development of automatic systems for sexist or offensive language detection.

1 Task Description

The task proposes the annotation of sexist and offensive content in the Romanian language from Twitter. Twitter is a public, online micro-blogging social media platform where users can post up to 280-character of content, known as tweets. The participants of this task are asked to mark the presence of sexist and offensive content by mapping labels to a given tweet.

The annotation process comprises the following steps:
1. analyse of the tweet's content
2. discern the topic

[1] acc.Merriam-Webster

3. identify the intention and the sentiment within it
4. classify the tweet based on a predetermined list of labels.

Specifically, once it is established the tweet does not contain any sexist elements, it is further analysed whether the content is offensive (*more on how to identify offensive language in section 3*). However, if the tweet contains sexist elements (*more on how to identify sexism elements in section 4*), the participants must further establish the type of sexism, namely identify if the sexist content is directly targeted to a person (*Sexist direct*), or describes a person (*Sexist descriptive*), or it is a description of a sexual discrimination experience (*Sexist reporting*), and label it accordingly. Figure 1 shows the steps and the corresponding label for each case described above (*more about labels in section 5*).

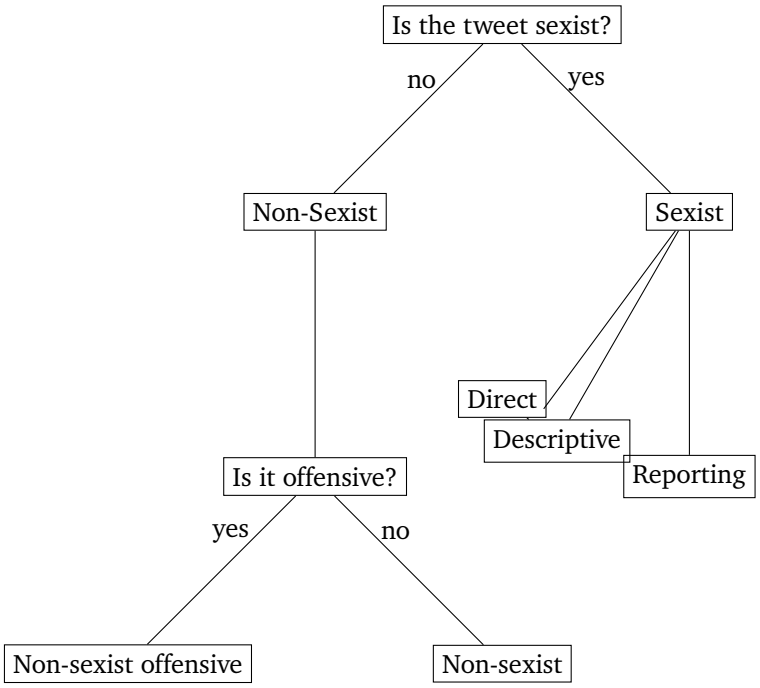

Figure 1: Data annotation workflow

2 Dataset Description

The data provided to the participants for the task comprises a subset of raw data set of tweets collected in real-time for a time period of 6 weeks, from 18.05.2021 until 01.07.2021. Generally, different elements can be found in a tweet:

(a) @Mentions: used to address another Twitter account;
(b) #Hashtags: categorizing content about a topic;
(c) Emoji or Emojis.

We encourage that all elements of a tweet to be also considered in the annotation process as they may provide more information about the topic, context, sentiment, and overall meaning of the tweet.

3 Offensive Content Description

We define offensive language to be cursing, profanity, blasphemy, epithets, obscenity, insults, aggression, hatred, racism, xenophobia, sexually explicit language, white nationalism, etc., except for sexism, which albeit is also part of the offensive language, it will be excluded from this category as for its marking we use different annotation labels.

E.g., "„Du-te-n mă-ta". → "Up yours"

4 Sexist Content Description

Sexism can appear in various forms on social media platforms. In order to define what constitutes to be sexist language, a subset of the collected tweets was analysed and together with a review of research papers (Chiril et al. (2020),Rodríguez-Sanchez et al. (2020), Elisabetta Fersini (2020), Basile et al. (2019), Hewitt et al. (2016), Hewitt et al. (2016), Anzovino et al. (2018)), we compiled a set of criteria suited for identifying sexist language, in a given tweet. Each category contains a set of examples in Romanian followed by their corresponding translation into English.

1. Objectification

E.g., "Om și femeie."
"The human being and the woman."
"Ce frumos, e copil sau e fetiță?"
"How beautiful, is it a child or a girl?"

The examples above present a mentality that objectifies the woman and categorizes her as a subspecies, namely a woman cannot be a human being and a girl cannot be a child.

2. Inferiority/ Incapacity

E.g., "Bărbatul este capul iar femeia gâtul."
"The man is the head and the woman the neck."
"La femei cică pe lângă CIP, le bagă și senzori de parcare."
"In addition to a chip, women will also get parking sensors."

Here, an archaic thinking is promoted, in which women are submissive to men, incapable of autonomy or personal power and are often discredited for their skills.

3. Role stereotyping

E.g.,"Locul femeii este la cratiță, iar rolul este sa facă copii și să-i crească."
"Women's place is in the kitchen, to birth children and raise them."

This belief eliminates any other purpose or autonomy for women.

E.g., "Când nevasta tace, să n-o întrerupi."
"*When the wife is silent, do not interrupt her.*"
"Soacră, soacră, poamă acră."
"*Mother-in-law, mother-in-law, you sour grapes.*"

Here, it is insinuated that wives are a nuisance, and mothers-in-law are overbearing. Either mother, wife, or mother-in-law, the role is stigmatized. However, obviously these labels can be assigned to both sexes.

E.g., "Trebuie sa te pui și tu la casa ta."
"You've got to settle down."
"Numai la măritiș ți-e capul."
"All you think about is getting married."
"Vei muri fată bătrână."
"You'll end up an old maid."

Here, there is an emphasis on a greater desire of women to marry and similarly, a greater pressure on them to get married. The marriage is the ultimate goals and if it is not met, the woman without a partner is viewed with suspicion or pity.

E.g., "Aia e femeie? Arată ca un bărbat."
"Is that a woman? She looks like a man. "

It shows that women's appearance needs to be in certain way.

4. Derailing

E.g., "Femeile care poartă fustă scurtă vor sa fie abuzate."
"*Women who wear mini skirts want to be abused.*"

To justify women abuse, sexually assaulted victims are partially blamed for the crime. This represents a derailing and absolves perpetrators from full responsibility.

5. Sexualization

E.g., "Ce bună ești!"
"You are sexy!"
"Esti o curvă."
"You slut."

Women are considered sexual objects and evaluated according to their physical characteristics and sexual character or they are addressed with gendered insults and slurs, most common used being *"bitch," "cunt," "slut," or "whore"*.

6. Hate and violence

E.g., "Urăsc femeile, ar fi mai bine fără ele."
"*I hate women, it would be better without them.*"
"Petrecerea se încinge cu maneaua și femeia cu cureaua."
"*The party heats up with manele*[2] *and the woman with the belt.*"
"Arată-mi sânii ăia frumoși."
"Show me those boobs."

The examples above show an intense dislike for women, aggressive pressure or intimidation, power assertion over women, advances of sexual nature, demands for sexual favors.

7. Male dominance

E.g., "Asta-i treabă de bărbat, de ce te bagi tu, nu înțeleg."
"*This is a man's job, why you do you interfere, I don't understand.*"
Gigi Becali: "Femeia nu poate fi președinte!"
Gigi Becali[3]*: "Women cannot be presidents!*"

Male domination is so ingrained in our consciousness that we don't even notice it. The above examples show an androcentric view in which women are simply excluded for certain positions or roles in society.

5 Corpus Labelling Conventions

In this project we follow a similar approach as (Chiril et al., 2020) for annotating sexist language, with a small difference, namely we further evaluate the non-sexist content, and check if it contains offensive language. We do not go in-depth in the types of offensive language, and therefore use only one class, namely *Non-sexist offensive*. The categories proposed are established according to the presence of sexist or offensive elements in a given tweet, as follows:

[2]genre of pop folk music from Romania.
[3]Romanian businessman and politician, mostly known for his ownership of a football club.

(a) **Non-sexist**

This label will be used if a tweet does not contain any sexist or offensive elements, or there are no sexist or offensive connotations. It may contain elements such as hashtags used in sexist texts, like #sexism #feminism #feminist #equality #racism #metoo #sexist #misandry #love #women #mensrights #misogyny #womenempowerment, #patriarchy #redpill #misandrist #gynocentrism#genderequality#womensrights #theredpill #mgtow #phenomenalact #lgbt #mensrightsactivist #maleissues #lgbtq #smashthepatriarchy #prochoice #bhfyp, #womensupportingwomen #doublestandards #siksilk #equalrights #mraemhumgg #fra #hra #n #gender #s #mensrightsactivism, #gynocentric k #toxicmasculinity, #feminazi #kad #discrimination #politics #girls #instagram #sexsy #india #z, however the content is neither sexist nor offensive.

E.g., "Campania MeToo, adică „și eu",a devenit o mișcare globală."
"The #MeToo campaign has become a global movement."

(b) **Non-sexist offensive**

This label will be used if a tweet does not have any sexist connotations, it does however contain offensive language.

E.g., "Ce bine va sta împreună, doi prosti amândoi."
"You two look good together, two fools."

(c) **Direct sexist**

This label will be used if a tweet contains sexist elements and it is addressed directly to a woman or group of women. We notice the presence of personal pronoun, 2nd person sg., pl., and imperatives, which indicates the addressee, as seen in the examples below:

E.g., "Tu ești femeie, de ce te bagi în discuții despre politică."
"You are a woman, why do you get involved in politics."
"Treci la bucătărie!"
"Go back to the kitchen!"
"Voi femeilor care va băgați ca musca-n lapte, tineți-vă gura."
"You women, that like to interfere, keep your mouth shut."

(d) **Descriptive sexist**

This label will be used if a tweet describes a woman or women in general. In descriptive sexist statements, women are not directly addressed, named entities (person, place, organization, product, etc.) are used instead. The addressing is done in the 3rd person, and the impact is less on the recipient compared to the addressing in the 2nd person, as it happens in the case of the *Direct sexist*.

E.g., "Locul femeilor este la bucătărie, toată lumea știe asta."
"Women's place is in the kitchen, everyone knows that."
Despre Viorica Dăncila : "Dacă apare o nouă bancnotă de 100 lei, va scrie pe ea una tută lei."
About Viorica Dăncila[4] *: "If a new 100 lei banknote is issued, one stupid lei will be written on it."*

(e) **Reporting sexist**

The label will be used if a tweet is a report of an experience or an act of sexism. The person reporting witnessed or heard from other sources, the act of sexism. The tweet can contain reporting verbs.

E.g., "Mi-a spus că e normal ca femeile însărcinate sunt plătite mai puțin pentru ca lucrează mai puțin."

"He told me that it is normal for pregnant women to be paid less because they work less."

"Ei s-au retras recent din Convenția prin care se stipula ca femeile primesc drepturi..."

"They recently withdrew from the Convention which stipulated that women receive rights... "

(f) **Cannot decide**

This label is dedicated for cases in which a tweet is not intelligible, it lacks context, and although it contains elements of sexism, the content is still ambiguous, or none of *c, d,* or *e* labels can be assigned.

Privacy policy

As far as it concerns the Twitter data sets utilised for this project, the content provided remains subject to the Twitter's Developer Agreement Policy, and must agree to the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy.

The purpose of the project

The **Corpus of sexist phrases in Romanian** is a project that has as the sole purpose of quantitative scientific research. To carry out this project, we annotate text, aimed at identifying sexist elements. Using the results of your answers, we build a corpus and conduct a quantitative study regarding the presence of sexism on social media platforms in Romania.

Personal Data Collection

We collect personal information from you, in general, information that could reasonably be used to identify and contact you and may include information such as:

- Contact details such as your name, email address, mobile phone number;
- Demographics, such as gender, professional status, and age.
- Survey data. When you fill out an annotation form, we collect your answers.

We do not collect financial information, data regarding racial or ethnic origin, political opinions, religious or philosophical beliefs. We may use your name for the purpose of assigning contributions to this project (mentioning your name in the *Acknowledgements* section of the scientific paper) if we have your prior consent.

How is your data collected?

We use two methods to collect data from and about you, by:

- Direct interactions
- You provide us with your email address when you fill out the annotation form.

References

M. Anzovino, E. Fersini, and P. Rosso. Automatic identification and classification of misogynistic language on twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems*, pages 57–64, Cham, 2018. Springer International Publishing. ISBN 978-3-319-91947-8.

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 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 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://www.aclweb.org/anthology/S19-2007.

P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully. 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 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.175.

P. R. Elisabetta Fersini, Debora Nozza. Ami @ evalita2020: Automatic misogyny identification. In V. Basile, D. Croce, M. Di Maro, and L. C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. CEUR.org, 2020.

S. Hewitt, T. Tiropanis, and C. Bokhove. The problem of identifying misogynist language on twitter (and other online social spaces). *Proceedings of the 8th ACM Conference on Web Science*, 2016.

L. V. Kremin. Sexist swearing and slurs: Responses to gender-directed insults. *LingUU*, 1:18–25, 2017.

Merriam-Webster. sexism. In *Merriam-Webster.com dictionary*. URL https://www.merriam-webster.com/dictionary/sexism.

F. J. Rodríguez-Sanchez, J. Carrillo-de Albornoz, and L. Plaza. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576, 2020. doi: 10.1109/ACCESS.2020.3042604. URL https://doi.org/10.1109/ACCESS.2020.3042604.