A Novel Corpus for Automated Sexism Identification on Social Media in Turkish

Lütfiye Seda Mut Altın and Horacio Saggion
Universitat Pompeu Fabra, Department of Information and Communication
Technologies, C/Tànger, 122, 08018, Barcelona, Spain
lutfiyeseda.mut01@estudiant.upf.edu, horacio.saggion@upf.edu

Abstract

In this paper, we present a novel dataset for the study of automated sexism identification and categorization on social media in Turkish. For this purpose, we have collected, following a well established methodology, a set of Tweets and YouTube comments. Relying on expert organizations in the area of gender equality, each text has been annotated based on a two-level labelling schema derived from previous research. Our resulting dataset consists of around 7,000 annotated instances useful for the study of expressions of sexism and misogyny on the Web. To the best of our knowledge, this is the first two-level manually annotated comprehensive Turkish dataset for sexism identification. In order to fuel research in this relevant area, we also present the result of our bench-marking experiments in the area of sexism identification in Turkish.

Keywords: Text Classification, Turkish, Gender Discrimination, Misogynistic Language, Sexist Language, Annotated Corpus

1. Introduction

Sexism is defined as "prejudice or discrimination based on sex; especially, discrimination against women".[1] Past research has shown that everyday sexism has a vast negative sociological and psychological impact on people: On the one hand, at the sociological level, it represents stereotypes including gender status beliefs which are associated to social hierarchies and leadership statuses (Ridgeway, 2001). On the other hand, there is a demonstrated negative impact on psychological well-being which affects self-esteem and leads to anxiety and depression (Swim et al., 2001; Feigt et al., 2022). There is an increasing interest in detecting and handling sexist speech, particularly on social media where anonymity and the sheer scale of propagated messages makes moderation highly difficult with existing manual moderation or filtering methods.

Research on sexism identification on social media has received considerable attention from the Natural Language Processing (NLP) community, with abundant research efforts in languages such as English, Spanish, and Italian (Kirk et al., 2023; Fersini et al., 2018; Rodríguez-Sánchez et al., 2021). However, low-resource languages (from the NLP perspective) such as Turkish have yet to be covered in this relevant domain. Our work aims to fill the existing gap in resources in the area of sexism identification in Turkish by releasing a new manually curated two-level dataset for the NLP community.

The rest of the paper is organized as follows: In section 2, we present an overview of the previous work in the field. In Section 3, we describe our methodology to create the dataset. In Section 4, we provide the results of our investigatory experiments on the dataset. Finally, in Section 5, we give a conclusion and an outline for future work.

2. Related Work

With the ubiquitous presence of social media platforms in modern societies, the amount of content published over the years has exponentially increased. In a free-speech digital world moderation is of paramount importance, this is why detecting hate speech has taken a central stage in many social media platforms and news organizations, and automated tools for its identification are nowadays prominent. However, research that focuses on specific types of hate speech such as gender discrimination is still rather limited. One of the earliest works in the field (Hewitt et al., 2016) proposed a Twitter dataset where tweets were classified according to the presence of misogynistic language as a form of abuse. A finer grain collection of tweets was later on proposed by (Anzovino et al., 2018) with annotations in classes indicating (i) Discredit, (ii) Stereotype and Objectification, (iii) Sexual Harassment and Threats of Violence, (iv) Dominance, and (v) Derailing. The AMI Automatic Misogyny Identification shared task, for Italian and English Tweets (Fersini et al., 2018) included misogyny identification and categorization as objectives. The 2020 edition of AMI also proposed an analysis of the models' fairness in classification (Fersini et al., 2020). For French, we highlight the Twitter dataset created by (Chiril et al., 2020) which follows a two-level annotation schema while, for

[1] https://www.merriam-webster.com/dictionary/sexism

SIGUL2024 Workshop, pages 10–15

Arabic, the misogyny multi-label annotated dataset by (Mulki and Ghanem, 2021). We base our annotation schema on the sEXism Identification in Social neTworks (EXIST) shared task which covers Spanish and English languages (Rodríguez-Sánchez et al., 2021).

With respect to Turkish datasets in the field, we have identified the resource by (Çöltekin, 2020) on abusive Turkish comments and the hate speech dataset by (Beyhan et al., 2022). Moreover, (Toraman et al., 2022)'s dataset relates to offensive gender topics and classifies them as *hate, offensive or normal*. Our analysis indicates that none of these Turkish datasets are solely focused on sexism identification and categorization.

3. Dataset

Given the lack of resources in the area for Turkish, we have created the first dataset on sexism identification following a process of annotation schema definition, data collection, expert annotation, and consolidation.

3.1. Data Collection

Following an already established methodology, data collection was carried out on X (formerly Twitter) and Youtube using their respective APIs[2] by issuing several focused queries. For YouTube, popular music videos have been selected from which we have extracted comments under the videos. To gather tweets from Twitter API, generic query exclusion criteria have been defined such as excluding re-tweets or tweets including images and videos. Queries were limited to the Turkish language with emojis kept, as they might carry valuable information. In addition, since Twitter Search API was normalizing special Turkish characters (ğ, Ğ, ç, Ç, ş, Ş, ü, Ü, ö, Ö, ı, İ), careful selection of keywords was considered so as to discard words would mean something different if normalized (e.g. 'şık' in Turkish means 'chic' whereas the normalized version 's*k' is a profane word.).

Queries were formed as a set of keywords and hashtags identified as potentially falling under one of the sexism categories, such as profane words indicating sexual violence. Keywords selection was based on various methods including not only common sense but also dictionaries created for gender equality or offensive terminology, certain viral events which may trigger inappropriate commentaries and additional keywords from initial test queries that returned sexist comments. As an example, a recent viral debate centered around the repeal of the legal regulation known as the Istanbul Convention, which addresses domestic violence was chosen as it contained misogynistic comments. In addition, time plays an important role in text classification since particular topics may only occur in specific time-period, dates were also considered for the searches. The full list of search keywords is provided along with the dataset.

3.2. Classification Schema

As the classification and categorization schema, EXIST 2021: sEXism Identification in Social neTworks classification was taken as the base reference and after some sample annotation trials, some minor modifications were done in the terminology and the definitions to adapt to cultural nuances (Rodríguez-Sánchez et al., 2021). At initial annotation trials, annotators labeled instances containing any statement related to politics (e.g., the name of a politician) as ideological, regardless of whether the instance included any sexism. To provide more clarity, a new terminology was introduced, for discrediting of the feminist movement as 'anti-feminism'.

- **Sexism Identification:** Level 1 class is defined as '**Sexist**' or '**Not-Sexist**'. Therefore, anything that does not include concepts in the sexism definition is classified as 'Not-Sexist'.

- **Sexism Categorization**: Sexism is classified in different categories. Definitions are based on EXIST 2021 with minor modifications. See Table 1 for examples of each type (in both the original language Turkish (TR) and English (ENG)).

Stereotyping, ideological thinking or dominance: The text expresses false ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving etc), or claims that men are somehow superior to women.

Objectification: The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles (compliance with beauty standards, hyper sexualization of female attributes, women's bodies at the disposal of men, etc.).

Misogyny and non-sexual violence / hatred towards women: The text expresses hatred and violence towards women.

Obscenity or Sexual violence: Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made. The examples in this category usually include the highest level of profanity.

Anti - Feminism: The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.

[2] Note that our collection was carried out before the restriction imposed by Twitter in recent months.

Category	Example
Stereotyping	TR: *@USER kadinlar mumkunse futbol konusmasin* (*@USER women should not talk about football if possible*)
Objectification	TR: *@USER Ablada tam ** +30 yaş üstü,evde kalmış kadın** sendromu var sanki. Gereksiz asabiyet,kendisini çevresinden farklı sanması ve hayvanlı foto.* (*It's like she has the ** 30+ year old, spinster woman ** syndrome. Unnecessary irritability, feeling different from her surroundings and a photo with animals.*)
Misogyny	TR: *@USER Kadınların beyni yok* (*@USER Women have no brains*)
Obscenity	TR: *@USER S*X S*X. Tecavuz den kacinamazsan zevk alacaksin, hala anlayamadinmi ahmakk* (*"@USER S*X S*X. If you can't avoid rape, you will enjoy it, haven't you figured it out yet, idiot*)
Anti feminism	TR: *@USER ... Erkekleşmiş, feminist kadın kılığında, kadinlıktan çıkmış, kadinlardan uzak durun.* (*"@USER ... Behind every successful man there is a woman.. Stay away from women who have become masculinized, disguised as feminist and unfeminine.*)

Table 1: Level-2 annotations for tweets in the dataset

Class	**# instances**	**% instances**
Not-Sexist	3167	45.8
Sexist	3748	54.2
Sexual Violence	1352	19.6
Stereotyping	1124	16.3
Misogyny	655	9.5
Objectification	468	6.8
Anti-Feminism	149	2.2
TOTAL	6915	100

Table 2: Dataset instances by category

3.3. Data Annotation

Based on EXIST (Rodríguez-Sánchez et al., 2021), a data labeling guideline was adapted to our data and refined after an annotation trial. Since annotation on current annotation platforms such as Amazon Mechanical Turk (AMT) demonstrated to be rather ineffective, we hired the services of a non-profit organization called "SistersLab"[3] that works for gender equality in the STEM fields. Their volunteers and experts, native Turkish speakers and involved in gender studies or volunteer actively in the field of gender equality, were engaged for the annotation process. For each instance in our dataset (Tweet or Youtube Comment) at least 2 agreed annotations have been requested for the target schema. In case there was no agreement between the first and second annotation a third annotation was requested.

Our final dataset resulted in **6,915 instances** of which **54.2%** is annotated as some type of 'sexist' content. See 2 for the distribution of categories in the dataset. **Cohen's Kappa** (Cohen, 1960) was used to calculate inter-annotator agreement and resulted in a value of **0.68** for **Level 1** which refers to **substantial** agreement; and a value of **0.55** for **Level 2** which refers to **moderate** agreement. Lower inter-annotator agreement for Level 2 than Level 1 annotation is expected due to the variety and complexity of the sub-types. Moreover, as some text might include more than one sub-type, even though the annotators have been advised to choose the most dominant type it added more complexity to the annotation process.

4. Preliminary Experiments

We have carried out a set of initial experiments in order to evaluate the dataset for comment classification experiments. Level 1 (sexism identification) was used for binary classification while Level 2 (sexism categorization) was used for multi-class classification. F1-scores were used to assess model performance. For the experiments described below we applied a fixed train (90%) and test (10%) partitions. Initially, we have tried a Support Vector Machine (**SVM**) (Cortes and Vapnik, 1995) model (linear kernel, C = 0.1, gamma='auto') training on a Term Frequency - Inverse Document Frequency (**TF-IDF**) vectorization. We have also used a neural network architecture **bi-LSTM** feed with word embedding. More concretely, the model consists of a **word embedding** layer (embedding dimension=300) implemented with a FastText (Joulin et al., 2016) model for Turkish ('cc.tr.300.bin') and a bidirectional Long Short-Term Memory (**bi-LSTM**) (Graves and Schmidhuber, 2005) layer (epochs = 10, batch size = 32) which

[3] https://sisterslab.org/

Model	Level-1	Level-2
SVM	0.88	0.70
bi-LSTM	0.89	0.70
BERT (multilingual)	0.87	0.72
BERT (Turkish)	0.87	0.73

Table 3: Classification results with neural models

is capable of capturing contextual information in both forward and backward directions.

Further experiments were run using Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018). The model used was extracted from Simple Transformers NLP (Wolf et al., 2020) library by HuggingFace [4].
A version of BERT for the Turkish language, **BERTurk** (bert-base-turkish-cased)[5] which is a community driven BERT model, trained on various Turkish corpora was used. And, for comparison, we also applied the **multilingual BERT** (bert-base-multilingual-cased)[6] which is pretrained on the top 104 languages of Wikipedia.
In Table 3 we show results for experiments involving neural networks which were trained with 90% of the training data and evaluated with 10% of test data.
In Figure 1, we present the confusion matrix for Level 1 classification (sexist / not-sexist) based on the predictions of the SVM model which has 90% train / 10% test data-set split.

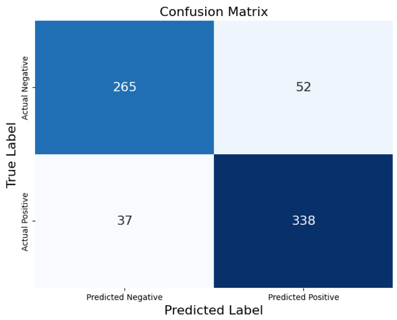

Figure 1: Binary classification - confusion matrix

In Figure 2, we present the confusion matrix for Level 2 classification. Number representations of labels corresponds to the following classes: 0:Not-Sexist, 1:Streotyping, 2:Anti-Feminism, 3:Misogyny, 4:Sexual Violence, 5:Objectification.
A manual **error analysis** was also done based on false predictions corresponding to the SVM model. Some of the findings and examples are as follows:

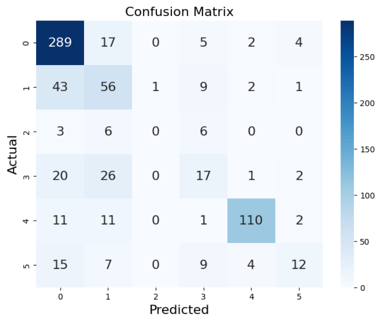

Figure 2: Multi-class classification - confusion matrix

Opposing opinion to the hate speech: The writer actually **opposes** to the sexist speech including the sexist speech in the sentence. The below example is predicted as 'Sexist'/'Stereotyping'; however it is actually 'Not-Sexist'. In addition, the writer uses this punctuation '(!)' to express irony.

> (ENG) @USER They want her to give birth to children and stay at home, not to work or study like a man (!) "Break your knees and sit at home", that's what they want.
> (TR) @USER İstiyorlar ki çocuk dogurup evde otursun,erkek gibi(!) çalışmasın,okumasın."Kır dizini otur evinde" istedikleri bu.

Idiomatic expressions with sexist background such as the example below is falsely labeled as not-sexist whereas its actual label is sexist. This example is **sentimentally quite positive** and not a hate speech directed to women; however the impression 'like a man' itself is a sexist idiom.

> (ENG): @USER You love like a man, my friend
> (TR): @USER Adam gibi seviyorsun kankam

5. Conclusion and Future Work

We created a manually annotated corpus for Sexism identification in Turkish on social media. Our corpus consists of 6915 instances which 54% of them contains a type of sexism. The dataset is publicly available to the research community [7]. To the best of our knowledge, this is first comprehensive dataset focusing sexism identification in Turkish. For future work, we would like to execute further Turkish specific pre-processing, data augmentation with language generation models and training on ensemble models.

[4] https://huggingface.co/transformers/

[5] https://huggingface.co/dbmdz/bert-base-turkish-cased

[6] https://huggingface.co/bert-base-multilingual-cased

[7] https://github.com/smut20/Turkish_Sexism_Dataset

Acknowledgments

We acknowledge support from the Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MICIU/AEI /10.13039/50110001103 and with the support from Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021).

6. Bibliographical References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A turkish hate speech dataset and detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Nicole D Feigt, Melanie M Domenech Rodríguez, and Alejandro L Vázquez. 2022. The impact of gender-based microaggressions and internalized sexism on mental health outcomes: A mother–daughter study. *Family Relations*, 71(1):201–219.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *CEUR workshop proceedings*, volume 2263, pages 1–9. CEUR-WS.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2020. Ami@ evalita2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Habibe Karayiğit, Ali Akdagli, and Çiğdem İnan Aci. 2022. Homophobic and hate speech detection using multilingual-bert model on turkish social media. *Information Technology and Control*, 51(2):356–375.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Hala Mulki and Bilal Ghanem. 2021. Let-mi: an arabic levantine twitter dataset for misogynistic language.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of exist 2023: sexism identification in social networks. In *European Conference on Information Retrieval*, pages 593–599. Springer.

Cecilia L Ridgeway. 2001. Gender, status, and leadership. *Journal of Social issues*, 57(4):637–655.

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, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69:229–240.

Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23(4):1159–1164.

Janet K Swim, Lauri L Hyers, Laurie L Cohen, and Melissa J Ferguson. 2001. Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social issues*, 57(1):31–53.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer.

Yasmen Wahba, Nazim Madhavji, and John Steinbacher. 2022. A comparison of svm against pretrained language models (plms) for text classification tasks. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 304–313. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.