Inters8: A Corpus to Study Misogyny and Intersectionality on Twitter

Ivan Spada[1], Mirko Lai[1,2] and Viviana Patti[1]

[1]Computer Science Department - University of Turin, Italy
[2]Aequa-tech srl, Turin, Italy

Abstract
This paper presents our research on the detection of online misogyny on social media and its intersection with other hate categories. Focusing on the phenomenon of misogyny, we carried out a corpus-based data analysis around victims of online hate campaigns. Targets were selected to study how misogyny and sexism intersect with other categories of social hatred and discrimination such as xenophobia, racism, and Islamophobia. This study includes an event-driven analysis of hate on Twitter concerning specific targets, the process of developing the *Inters8* corpus, and its manual annotation according to a novel multi-level scheme designed to assess the presence of intersectional hatred.

Keywords
hate speech, automatic misogyny identification, intersectionality, annotated corpora, social media

***Warning**: This paper contains examples of potentially offensive content.*

1. Introduction

The term *intersectionality* refers to the coexistence of multiple forms of social categorizations such as ethnicity, gender, sexual orientation, social class, disability, etc., which can lead to discrimination and generate obstacles in the daily lives of those affected [1]. In the specific case of misogyny, discrimination takes different shapes depending on the other co-existing forms of discrimination such as racism, classism, ableism, or homophobia [1]. Therefore, the phenomenon should not be studied in isolation.

The coexistence of different forms of discrimination suggests the study of social interactions and the intersectionality between multiple categories of hate. It is interesting to explore how language may vary when interactions involve people who are at the intersection of multiple discriminated social categories that are henceforth referred to as *dimensionalities.* In particular, [1] explained how intersectionality between multiple dimensionalities may generate a new discriminated category framed and treated differently.

The objective of this paper is to analyze how misogynistic hatred intersects with other dimensionalities and how this appears in a micro-blogging social platform such as Twitter[1]. The first contribution of this paper is an analysis of gender discrimination, inequality comparisons, presence of stereotypes, also oriented to shed light on how users interact with, support, and attack targets of misogynous hatred on social media. As a second contribution, we created *Inters8*, an Italian corpus containing a subset of TWITA [2] filtered by following a target and event selection process, for the purpose of studying intersectionality. Thereafter, a portion of the corpus related to Silvia Romano's liberation [3] was annotated (called *Inters8_SRomano*) so as to explore characteristics of intersectional hatred in a specific pilot study.

Outine. After a brief technical contextualization (Section 2), we describe the target-event driven analysis, the creation of the *Inters8* corpus and the annotation task (Section 3). Then, the novel annotation scheme applied to the data is described (Section 4) and a discussion of the outcome of the annotation process is presented (Section 5). Conclusions and future work end the paper.

2. Related work

Twitter, as well as other online social media platforms, employs algorithms for detecting content that violates its terms and conditions committed to making Twitter a safe place for users. As observed in [4], the relationship between ethnicity and gender can influence the recognition of false positive hateful content of African Americans, especially in the case of female users. Indeed, bias can permeate the system and reinforce AI discrimination.

Subjective phenomena that could be affected by social and cultural context such as misogyny, stereotype and racism, have to be approached recognizing individuals'

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

ivan.spada@unito.it (I. Spada); mirko.lai@unito.it (M. Lai); viviana.patti@unito.it (V. Patti)
0009-0002-0459-1189 (I. Spada); 0000-0003-1042-0861 (M. Lai); 0000-0001-5991-370X (V. Patti)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://twitter.com

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

perceptions that can greatly vary according to personal experiences and cultural backgrounds.

Several efforts in terms of automatic detection have been provided by scholars for countering hate speech [5, 6]. About misogyny, a first computational effort for the detection of misogyny in English tweets has been provided in [7], while [8] attempts to address the problem of measuring and mitigating unintended bias in machine learning models trained for misogyny detection. A first automatic Misogyny Identification (AMI) shared task has been organized within EVALITA and IberEval 2018 evaluation campaigns [8, 9] to detect misogyny in tweets in various languages (English, Spanish, and Italian). In particular, participants were specifically requested to identify if the message is misogynistic, and then to categorize the target (person or not) and the type of misogyny using the categories developed by [10]. Based on the availability of multilingual datasets targeting misogyny and other kind of abusive language, in [11] a multilingual and cross-domain study on misogyny identification in Twitter is proposed, where some insights on features of misogyny and on the interaction between misogyny and related phenomena are provided. In this work, we try to shed more light on misogyny and intersectionality with other co-existing forms of discrimination such as xenophobia, islamophobia, and stereotype proposing a new annotation scheme. In Section 4.1 we specifically analyze the contributions that inspired our work.

3. Methodology

In this section, we describe the methodological pipeline we designed in order to collect data to analyze intersectional hate and discrimination. The target- and event-oriented nature of hate speech in social media has been the object of recent studies [12, 13, 11, 14, 15]. Hate discourses may vary in relation to events and victims belonging to multiple dimensionalities. However, it is necessary to take into account that recognition of the discriminatory phenomenon may be partly subjective and influenced by the social and cultural context.

In order to contextualize and explore the phenomenon of intersectionality among multiple social categories subjected to hate and discrimination, we conducted an analysis of discourses concerning public people, known to Italian society, selected specifically for this task.

Our pipeline consists of a sequence of steps (Figure 1).

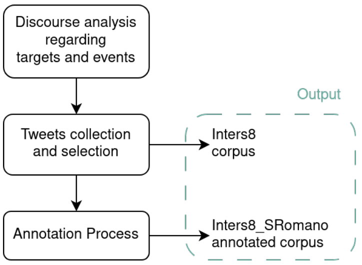
Discourse analysis regarding targets and events
Tweets collection and selection
Annotation Process
Output
Inters8 corpus
Inters8_SRomano annotated corpus

Figure 1: Pipeline for the creation and annotation of Inters8.

3.1. Discourse analysis regarding targets and events

The period chosen for analysis is the first half of 2020, the year when the COVID-19[2] pandemic began and the population forced to stay at home was making substantial use of social media [16]. The analysis process starts with prior knowledge of the Italian context obtained from consulting services that convey information such as newspapers, news broadcasts, and TV shows. It was noticed that some streams of discourse on Twitter were event-driven, these included: reporting news, inviting guests, and discussing known people on TV shows.

Focusing on misogyny (the explicit and implicit attitude of generic aversion to women), the additional social categories considered are inspired by those in Vox's Intolerance Map n.7[3]. The categories taken into account in this study are as follows: *misogyny*, *xenophobia*, *anti-semitism*, *Islamophobia*, *homolesbobitransphobia*, *political opinions* and *physical appearance*.

Through scraping news and TV shows available on RaiPlay[4] (the Italian television streaming service), we viewed episodes of 23 TV shows, qualitatively analyzed and selected 17 well-known people in Italy[5] who fall into multiple dimensionalities considered in this case study. The manual analysis of the language, expressed in TV programs and Twitter interactions concerning the selected targets, allowed the extraction of information framing the subjects: dimensionalities, topics, events, debates, hashtags, and the most common and narrow keywords. The following were annotated for each victim: Twitter username (if they joined the social network), characteristics potentially exposed to hate and discrimination, hashtags, time period analyzed, TV shows, and links to episodes where they were invited or talked about.

In addition, an NxM table[6] was compiled to make the target comparison visually easier, where N refers to the

[2] https://www.who.int/health-topics/coronavirus

[3] VOX, University of Milan, Sapienza University of Rome, Aldo Moro University of Bari and ItsTime (2022). The map of intolerance n.7. Available at: http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-7/

[4] https://www.raiplay.it/

[5] Giovanna Botteri, Carola Rachete, Cathy La Torre, Cécile Kyenge, Chiara Appendino, Diletta Leotta, Emma Bonino, Greta Thunberg, Ilaria Cucchi, Laura Boldrini, Liliana Segre, Michela Murgia, Rula Jebreal, Silvia Romano, Teresa Bellanova, Virginia Raggi and Vladimir Luxuria.

[6] https://github.com/ivsnp/inters8/tree/main/targets

people and M to the hate categories. Since misogyny was the focus of this study, all selected victims had the *misogyny* column marked. Other categories were marked when present.

Output: selection of related targets, events, and hashtags useful for the next phase.

3.2. Tweets collection and selection

Data Collection Given the targets and events obtained from the previous phase, Italian tweets regarding targets in the temporal surroundings of the detected events were extracted from TWITA [2]. These tweets contained at least one hashtag among those found during the first phase or both the first and last names of the corresponding target or aliases.

We collected the following metadata for each Twitter interaction: *tweetId*, *date*, *text*, *type* (tweet, retweet, quote, or reply).

Output: collection of tweets related to targets and events useful for the next phase.

Data Selection After obtaining the collection of tweets in output from the previous phase, we set out to create a corpus containing tweets related to events concerning targets that were potentially victims of intersectional hatred. This decision aimed to provide a set of Twitter interactions (tweets, replies, quotes, and retweets) to perform an analysis of the intersection of various dimensionalities in a target-event context.

We proceeded with target-event filtering in order to obtain a case study on which to start analyzing the phenomenon. The *Inters8* corpus was populated with the Twitter interactions collected by selecting the deliverance from captivity and homecoming of the target Silvia Romano[7] to Italy on May 9-10, 2020. This choice of target-event pair was made because there were more Twitter interactions compared with others [17][8]. Thus, it was intended to create a pilot on a specific case study. Given this target-event choice, the goal was to explore the intersectionality between the following dimensionalities: *misogyny*, *xenophobia*, and *Islamophobia*.

Inters8 contains 248240 interactions concerning the chosen target-event pair distributed during May 9-24, 2020. It consists of contents distributed per interaction type as follows: retweet 75%, tweets 18%, reply 4%, and quote 2%. The metadata is as follows: *tweetId* and *type* (tweet, retweet, quote, or reply).

Output: *Inters8* corpus consisting of the tweets collected regarding the selected target and event.

Sampling of Data to Annotate The process of creating the Inters8 subset followed these steps: (1) retweets removal, (2) similar tweets removal using *cosine similarity* by setting the threshold to 0.7, (3) the collection was filtered to include tweets with and without the Italian flag emoji with 50% proportion keeping the same distribution for days and hours. The latter decision was made so that the annotation could be compared according to the presence of the Italian flag emoji, days, and hours. Indeed, the presence of the Italian flag seemed to convey hateful content in the pilot study. The collection related to Silvia Romano and the selected event reached 3006 contents, 1500 were randomly extracted for creating a sample to be annotated *Inters8_SRomano*.

The subset was cleaned of user mentions and URLs. The metadata used to describe the tweets were as follows: *id*, *parentTweetText* (if exists), and *tweetText*.

Output: *Inters8_SRomano* annotated according to the proposed annotation schema

3.3. Annotation process

The manual annotation process was divided into two phases: (1) *pilot* - a sampling of 50 tweets was selected to evaluate the annotation scheme and (2) *operational* - the annotation of the *Inters8_SRomano* sample dataset.

Our annotation scheme is described in Section 4. Twelve annotators, balanced by gender, were employed in order to ensure a diversified and representative group covering a wide range of perspectives and experiences. Guidelines provided to annotators were refined after a discussion within the pilot phase.

The subset was annotated as follows: we collected 3 independent annotations for 1373 tweets. The remaining 127 were annotated only by two independent annotators; a third annotation was collected in order to solve the disagreement.

The quantitative analysis conducted by manual annotation enabled the assessment of the actual coexistence of multiple discriminatory and hate dimensions and the construction of a gold standard.

Annotators features table, annotation scheme, guidelines in Italian and English, and annotation results obtained through majority vote are available here.

Output: subset of *Inters8* annotated according to the proposed annotation scheme.

4. A Novel Annotation Scheme

This section describes the process of creating a novel annotation scheme for multi-level analysis of intersec-

[7] https://www.bbc.com/news/world-africa-52608614

[8] According to the Italian observatory "Map of Intolerance http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-5/ the peak of social attacks against Muslims occurred right around the time of Silvia Romano's liberation, a media shitstorm that prompted the Special Operations Group (Ros) to open up an investigation ad hoc into the matter.

tionality.

4.1. Related annotation schemes

Our annotation scheme is partially inspired by the ones designed in [18, 19]. The scheme used in the AMI@EVALITA18 shared task challenged the participants not only to determine whether misogynous content was expressed in the tweets, but also to classify the misogynistic behavior, by proposing the categories: *Stereotype & Objectification*, *Dominance*, *Derailing*, *Sexual Harassment & Threats of Violence*, and *Discredit*. A deeper analysis presented in [20], suggested some insights and motivations to simplify the fine-grained misogynistic behavior to be annotated in our scheme.

The distinction between specific individuals and generic groups of people, also mentioned in [19], was not introduced in our scheme, since the debates around the selected victims turn out to be particularly specific.

A further contribution, annotating an Italian immigration corpus, measured the intensity of hate speech on a scale of 0 to 4 [21]. The idea of measuring hate inspired the comparison of intensities and prevalence among coexisting dimensionalities.

Stance analysis in [22] is performed to check the behavior of tweets in response to others. They used the following labels: *agree-accept* (support), *reject* (deny), *info-request* (question), and *opinion* (comment). In the case under analysis, it was sufficient to consider: *support*, *against*, and *neutral*.

Moreover, as highlighted also in [19, 21], it is important to differentiate aggressive language from hate speech: in fact, aggressive content is not necessarily expressed through hateful vocabulary and *vice versa*. Since hatred and discrimination can appear implicitly, it is not always easy and immediate to recognize them in negative and aggressive content on social media. Moreover, not all expressions of disapproval and disagreement with groups imply discrimination. Such findings were useful in order to highlight the importance of annotating both implicit and explicit forms of hate speech.

4.2. Annotation scheme

The multi-level scheme[9] was meant to bring out the dimensionalities, and the cohesiveness and prevalence among them. The coarse-grained level is intended to annotate the presence of misogyny, xenophobia, and Islamophobia. The fine-grained analysis, first, aims to recognize which dimensionality prevails over the others. Secondly, a sub-classification of misogyny, if there is any, is proposed to classify it into *sexual harassment / derailing* and *discrediting / dominance*. Finally, an annotation of stereotyping, victim defense, and stance toward any parent tweet is proposed. A detailed description of the labels involved, supported by examples, follows.

Misogynistic behavior $[yes/no]$: explicit and implicit forms, including aversion, repulsion, target silencing and instrumentalization of pregnancy, following the definition "*misogynistic behavior is about hostility towards women who violate patriarchal norms and expectations, who aren't serving male interests in the ways they're expected to. So there's this sense that women are doing something wrong: that they're morally objectionable or have a bad attitude or they're abrasive or shrill or too pushy*" [23]. Annotate the two following sub-labels only if $misogyny = yes$:

- **Sexual harassment and/or derailing** $[yes/no]$: the first includes avance, requests for sexual favors, and any form of harassment involving sex or speech in which abuse of women is justified by belittling or evading male responsibility. The latter refers to the intention to divert support toward the victim by directing the discourse to a more comfortable alternative issue while ignoring the discriminatory problem.

  ITA: *La z****la, appassionata ai c**zi talebani, ha orchestrato una messinscena con il tipo che se la s**pa e si è sistemata a vita con il riscatto.*

  ENG: *That s**t, fond of Taliban c***s, orchestrated a setup with the guy who fu**ed her and set herself up for life with the ransom money*

- **Discredit and/or dominance** $[yes/no]$: discrediting occurs when an individual S, through a communicative act, damages the image of another individual T in front of a third party (individual or group A) by referring to actions or characteristics of T that are considered negative by A. Dominance is typically expressed as an assertion of superiority by highlighting gender inequality.

  ITA: *Conte dacci le prove del riscatto pagato dagli italiani per questa odiosa nullità e vergogna nazionale! È una bambina indottrinata, senza cervello e stupida. È andata in terre cesso per seguire le sue idiozie apparentemente umanitarie.*

  ENG: *Conte, give us evidence of the ransom paid by the Italians for this odious nothingness and national disgrace! She is an indoctrinated brat, braindead and stupid. She went to toilet-lands to follow her supposedly humanitarian nonsense*

[9] Annotation scheme and guidelines are available at: https://github.com/ivsnp/inters8/blob/main/annotation

Xenophobia and/or racism $[yes/no]$: explicit and implicit forms, i.e., expressions of racism based on the arbitrary assumption of the existence of biologically and historically "superior" human races, aversion to foreigners, and what is foreign. The latter manifests itself in attitudes and actions of intolerance and hostility toward the culture and inhabitants of other countries. In order to analyze *ingroup* and *outgroup* dynamics [24, 25, 26], we also consider texts where the target subject and other Italians are insulted or rejected as members of the *ingroup*, or stigmatized as anti-Italian, because of their proximity to (or support for) foreign populations or immigrants, to be expressions of xenophobia.

> **ITA**: *Una donna bianca convertita all'Islam, esce indenne dai ne***ni, belve inferocite, andate tutti a fare in c*lo.*
>
> **ENG**: *A white woman converted to Islam, comes out untouched by the nig***s, raging beasts, f**k you all.*

Islamophobia $[yes/no]$: explicit and implicit forms of strong aversion, dictated by prejudicial reasons, toward Islamic culture and religion. The main manifestations include criminalizing targets by describing them as threatening and violent. It is often joined by xenophobia and may appear in the form of dehumanization of targets.

> **ITA**: *È venuta qui per fare attentati, è una terrorista, è anche incinta di un musulmano. Se stava bene in Islam rimpatriatela #convertita.*
>
> **ENG**: *She came here to do bomb attacks, she is a terrorist, she is also pregnant by a Muslim. If she was fine in Islam then send her back #converted*

Prevalence $[misogyny/xenophobia\ and/or\ racism/Islamophobia/absent]$: in case of coexistence of at least two of the main categories to be analyzed, indicate which one prevails over the others within the tweet.

> **ITA**: *Il governo ruba 4 milioni di euro agli italioti per pagare uno specie di riscatto al marito islamico che la mette incinta e la converte. Arriva in Italia contenta, ingrassata e viene accolta come una santa. Popolo idiota!*
>
> **ENG**: *The government steals 4 million euros from the Italians to pay some kind of ransom to her Islamic husband who impregnates her and converts her. She arrives in Italy happy, fat and is welcomed as a Saint. Idiotic people!*
>
> **Note**: *This tweet is an example of the intersection of multiple dimensionalities of hatred toward vulnerable groups. The term "italiota" means "Italian idiot" and "ingrassata" refers to the target's pregnancy.*

Stereotypes $[yes/no]$: negative sexist, xenophobic, racist and Islamophobic stereotypes concerning vulnerable groups targeted by discrimination and hate speech considered in this study on intersectional hatred. Stereotyping is a generalization conducted about a group of people, in which characteristics are attributed to all members of the group [27]. Stereotyping is based on a set of beliefs, not based on experience, that people enact to interpret their surroundings and move through them.

> **ITA**: *È venuta qui per fare attentati, è una terrorista, è anche incinta di un musulmano. Se stava bene in Islam rimpatriatela #convertita.*

Target defense $[yes/no]$: it indicates whether the user who posted the tweet defends the hate target, contributing to creating a counter-narrative effect. It includes both support without discrimination and support that redirects hatred toward other people (without actually counteracting hate speech).

> **ITA**: *Il privato di Silvia Romano non dovrebbe essere nel dibattito pubblico. È stata liberata da una prigione fisica ma intrappolata in una di violenze psicologiche e pregiudizi inutili e ingiusti.*
>
> **ENG**: *Silvia Romano's private life should not be in the public debate. She was released from a physical prison but trapped in one of psychological violence and unnecessary and unjust prejudice.*

Stance $[support/against/neutral/absent]$: it identifies the stance of the user who posted a tweet reacting to another one. The label takes the value *support* if it agrees with the parent tweet, *against* if it disputes it, and *neutral* if no stance can be inferred from the text. If there is no parent tweet the default value is *absent.*

> **ITA**
> **Parent tweet**: *La liberazione di Silvia Romano è una bella notizia. L'aspettiamo in Italia, ringraziamo i nostri servizi di Intelligence e coloro che hanno contribuito a questo importante obiettivo.*
>
> **Child, or reply, tweet (contestazione)**: *Assolutamente no! Vanno a fare le splendide in Africa ma quando si accorgono che*

*ci sono i ne***ni cattivi chiedono aiuto a mamma Italia.*

ENG:

Parent tweet: *The liberation of Silvia Romano is good news. We are waiting for her in Italy, we thank our intelligence services and those who contributed to this important achievement.*

Child, or reply, tweet (against): *Absolutely not! They go show off in Africa, but when they realise there are bad nig***s they ask Mamma Italia for help.*

5. Results

The annotation of *Inters8_SRomano* extracted from *Inters8* and the harmonization stage by majority vote yielded the results shown below.

Islamophobia is the most annotated dimensionality, followed by misogynistic behavior and xenophobia/racism.

Label annotation detected the following amounts of tweets in the subset (see Figure 2): *misogynistic behavior* 288 (19.2%), *sexual harassment and/or derailing* 36 (12.5% of misogynistic behavior), *discredit and/or dominance* 247 (85.8% of misogynistic behavior), *xenophobia and/or racism* 153 (10.2%), *Islamophobia* 317 (21.1%), *stereotype* 394 (26.3%), *target defence* 501 (33.4%), and *stance* [against=119, support=108, absent=42, neutral=24]([40.6%, 36.9%, 14.4%, 8.2%] out of 293 reply tweets).

Stance, on the other hand, appeared difficult to annotate because the stances often went off-topic.

Among tweets labeled with at least one of the three main dimensions of hate included in the proposed annotation scheme, the Italian flag (in *name*, *screen-name*, *bio* or *tweet*) appears as follows: 78.5% of misogynistic behavior, 73.2% of Xenophobia and/or racism and 74.4% of Islamophobia.

The subset of tweets annotated with an intersection between the three main dimensionalities (at least 2) has cardinality 222. The most present and prevalent dimension was Islamophobia.

The following label distributions were obtained from the annotation of intersectional tweets (see Figures 3 and 4): *misogynistic behavior* 184 (82.9%), *xenophobia and/or racism* 117 (52.7%), *Islamophobia* 199 (89.6%), and *prevalence* [Islamophobia=91, misogynistic behavior=59, absent=42, xenophobia and/or racism=30]([41%, 26.6%, 18.9%, 13.5%] out of 293 reply tweets).

5.1. Inter-Annotator Agreement

Cohen's average Kappa [28] was 0.40 and the Fleiss' Kappa calculated was as follows: misogyny 0.49, sexual

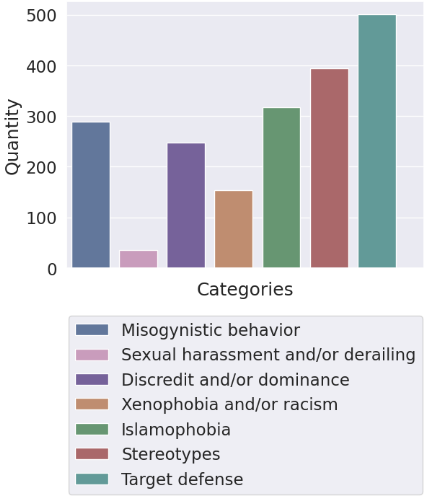

Figure 2: Annotation of binary categories.

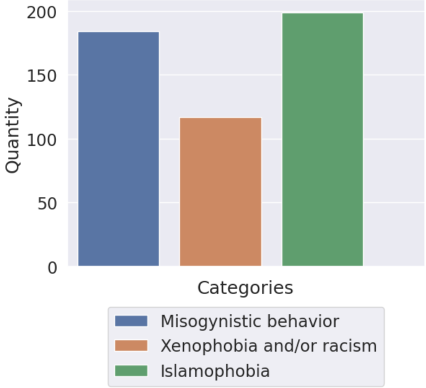

Figure 3: Three main dimensionalities annotated with the value YES in the intersection subset (at least two dimensionalities=yes).

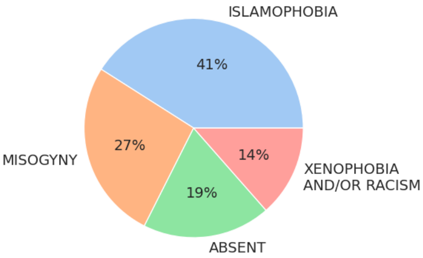

Figure 4: Annotation of prevalence among dimensionalities in subset intersection (at least two dimensionalities=yes).

harassment / derailing 0.29, discredit / dominance 0.44, xenophobia / racism 0.32, Islamophobia 0.53, prevalence 0.33, stereotyping 0.33, target defense 0.59 and stance 0.81. The calculation of Fleiss' Kappa for tweets with three independent annotations showed how complicated it was to classify the domain across the dimensionalities of the established multi-level scheme.

5.2. Considerations

Misogyny has been labeled more by female annotators, only $1/3$ of male annotators come close to the former. This is an expected result as the former tend to be more sensitive to the issue confirming that, for creating an unbiased annotated dataset, is important to employ annotators belonging to heterogeneous social categories.

Inter-Annotator Agreement Following an overview of the annotated subset and comparison with the annotators involved in this experiment, it appeared that many disagreements occurred on the label *stereotypes*, some annotators recognizing many more than others independently from their self-identified gender and age. The annotations appeared quite subjective and often repetitive because the presence of other discriminatory dimensionality often involves stereotypes. Concerning *stance*, the presence of irony and rhetorical questions inside the dataset complicates the valuation of the attribute.

The complexity of some annotation scheme labels emphasizes the difficulty of annotating tweets about this domain and brings out the presence of bias.

In addition, the phenomenon of *premediation* [29] has been observed in this case study. Indeed, Twitter's users expressed their own opinions favoring immediacy and emotionality in communication as a preliminary reaction to the first information about the news. Tweets and interactions began immediately, despite the fact that the full picture of the affair was not clear at that moment, bringing the event to the platform's trending topics.

Target Comparing the Twitter interactions regarding Silvia Romano with those around Luca Tacchetto[10], Alessandro Sandrini[11], and Sergio Zanotti[12], the following emerged. The three Italian people listed were victims of kidnapping like Silvia Romano. Among them, Tacchetto and Sandrini were converted to Islam. Considering the first week after the target subjects' homecoming, the following amounts of interactions were detected: Romano 237031, Tacchetto 1668, Sandrini 546, and Zanotti 2206. Interestingly, the volume of reactions related to the Silvia Romano's liberation is much higher (see [30] for a deeper analysis about this topic), suggesting that the intersectionality with misogyny matters.

Recurrent topics Among the tweets annotated, some recurring topics emerged: (1) dissent on the economic plan arguing that payment of the alleged ransom for release was not necessary, (2) aesthetic appearance by seeing physical appearance and clothing as objects of dialogue, (3) being ungrateful, selfish and a traitor to her (Silvia Romano) country for converting to the Islamic religion, (4) victimization, (5) pregnancy, and (6) politics.

In the former case, it was not always possible to recognize the discriminatory nature since a variable related to discontent about the Italian economic situation was also present. The second focused on arrival at the airport wearing the hijab and a watch pointed to as lavish despite the fact that it was not possible to distinguish it from the content disseminated by the media. The third appeared describing her as a member of the *outgroup*. The remaining ones also appeared frequently in the narrative of the annotated subset.

Counter-speech It has been observed that *counter-speech*, carried out by users who take the defense of victims, sometimes proposes an alternative narrative to hate speech. Other times they follow defensive strategies that are themselves offensive generating further hate speech.

6. Conclusions and future work

Developing the *Inters8* corpus built considering an intersectional target-event pair allowed us to explore a case study and analyze Twitter interactions related to Silvia Romano on social media. The manual annotation applied highlights how multiple dimensionalities coexist and intertwine in cases of intersectional hate.

Despite the evidence of the phenomenon and its dynamics, results presented here are related to the specific case study taken into account, and to the target and the event selected. In fact, at the current stage of development, the *Inters8* corpus includes content related to the specific intersectional Silvia Romano's liberation target-event pair. We plan to expand the corpus with additional targets, events, and social categories. It would then be interesting to compare multiple targets in the same intersection and study how local culture might influence the phenomenon over several countries.

As the corpus is built around the Italian context, the data are exclusively in Italian. The integration of multiple languages would allow for greater generalization and the study of the geographic distribution of the phenomenon around known people and events.

Finally, there are many interactions on social networks and the experimental study for automatic detection of intersectional hate may be a challenge of particular interest.

[10] https://www.nytimes.com/2020/03/14/world/africa/mali-hostages-released.html

[11] https://apnews.com/article/---0acfda0fe2974efab72081965cb7d3c6

[12] https://apnews.com/article/01e8c9c94f8e4441b7ad68de3d58e667

References

[1] K. W. Crenshaw, Mapping the margins: intersectionality, identity politics, and violence against women of color, Stanford Law Review 43 (1991) 1241–1299.

[2] V. Basile, M. Lai, M. Sanguinetti, Long-term social media data collection at the university of turin, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018, volume 2253 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–6. URL: https://ceur-ws.org/Vol-2253/paper48.pdf.

[3] L. Berlingozzi, “welcome home silvia, into the lion’s den": How gender biased narratives frame hostage liberations, Security Praxis (2020). URL: https://securitypraxis.eu/silvia-romano-gender-biased-narratives-hostages/.

[4] J. Kim, C. Ortiz, S. Nam, S. Santiago, V. Datta, Intersectional bias in hate speech and abusive language datasets, CoRR abs/2005.05921 (2020). URL: https://arxiv.org/abs/2005.05921. arXiv:2005.05921.

[5] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523. URL: https://doi.org/10.1007/s10579-020-09502-8. doi:10.1007/s10579-020-09502-8.

[6] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, PLOS ONE 15 (2021) 1–32. URL: https://doi.org/10.1371/journal.pone.0243300. doi:10.1371/journal.pone.0243300.

[7] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on Twitter, in: M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, F. Meziane (Eds.), 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, volume 10859 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 57–64. URL: https://doi.org/10.1007/978-3-319-91947-8_6.

[8] D. Nozza, C. Volpetti, E. Fersini, Unintended bias in misogyny detection, in: IEEE/WIC/ACM International Conference on Web Intelligence, WI ’19, Association for Computing Machinery, New York, NY, USA, 2019, p. 149–155. URL: https://doi.org/10.1145/3350546.3352512. doi:10.1145/3350546.3352512.

[9] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, volume 2150 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 214–228. URL: https://ceur-ws.org/Vol-2150/overview-AMI.pdf.

[10] B. Poland, Haters: Harassment, abuse, and violence online, U of Nebraska Press, 2016. URL: https://books.google.it/books?id=Jd4nDwAAQBAJ.

[11] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360. URL: https://www.sciencedirect.com/science/article/pii/S0306457320308554. doi:https://doi.org/10.1016/j.ipm.2020.102360.

[12] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 907–914. URL: https://aclanthology.org/2021.acl-short.114. doi:10.18653/v1/2021.acl-short.114.

[13] E. W. Pamungkas, V. Patti, Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370. URL: https://aclanthology.org/P19-2051. doi:10.18653/v1/P19-2051.

[14] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, V. Patti, Emotionally informed hate speech detection: A multi-target perspective, Cognitive Computation 14 (2022) 322–352. URL: https://doi.org/10.1007/s12559-021-09862-5. doi:10.1007/s12559-021-09862-5.

[15] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of your hate: The challenge of time in hate speech detection on social media, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/12/4180. doi:10.3390/app10124180.

[16] J. Brailovskaia, J. Margraf, Addictive social media use during covid-19 outbreak: Validation of the bergen social media addiction scale (bsmas) and investigation of protective factors in nine countries, Current Psychology (2022). doi:10.1007/s12144-022-03182-z.

[17] C. Annovi, G. Dentice, F. Manenti, F. Portoghese, V. Lazzerini, Y. Pallavicini, V. Gullo, Le cause di

discriminazione, hate speech e crimini d'odio contro le donne musulmane in Italia, Technical Report, Deliverable 2.1, project "TRUST: Tackling Under-Reporting and Under-Recording of Hate Speech and Hate Crimes Against Muslim Women", co-funded by European Union, Grant Agreement no. 101049611, 2022. URL: https://www.trust-project-eu.info/trust/wp-content/uploads/2023/04/TRUST-D2.1-ITA-1.pdf.

[18] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (AMI), in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–9. URL: https://ceur-ws.org/Vol-2263/paper009.pdf.

[19] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[20] S. Lazzardi, V. Patti, P. Rosso, Categorizing misogynistic behaviours in italian, english and spanish tweets, Proces. del Leng. Natural 66 (2021) 65–76. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6323.

[21] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. A. Stranisci, An italian twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 2798–2805. URL: https://aclanthology.org/L18-1443".

[22] E. W. Pamungkas, V. Basile, V. Patti, Stance classification for rumour analysis in twitter: Exploiting affective information and conversation structure, in: A. Cuzzocrea, F. Bonchi, D. Gunopulos (Eds.), Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018, volume 2482 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–7. URL: https://ceur-ws.org/Vol-2482/paper37.pdf.

[23] K. Manne, Down Girl: The Logic of Misogyny, Oxford University Press, 2017. URL: https://doi.org/10.1093/oso/9780190604981.001.0001. doi:10.1093/oso/9780190604981.001.0001.

[24] G. Comandini, V. Patti, An impossible dialogue! nominal utterances and populist rhetoric in an Italian Twitter corpus of hate speech against immigrants, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 163–171. URL: https://aclanthology.org/W19-3518. doi:10.18653/v1/W19-3518.

[25] B. Sauer, A. Krasteva, A. Saarinen, Post-democracy, party politics and right-wing populist communication, Routledge, 2018, pp. 14–35.

[26] G. Mazzoleni, R. Bracciale, Socially mediated populism: the communicative strategies of political leaders on facebook, Palgrave Communications 4 (2018) 1–10. URL: https://EconPapers.repec.org/RePEc:pal:palcom:v:4:y:2018:i:1:d:10.1057_s41599-018-0104-x.

[27] E. Aronson, T. Wilson, R. Akert, Social Psychology, Always Learning series, Pearson, 2013. URL: https://books.google.it/books?id=wr9uvgAACAAJ.

[28] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37 – 46.

[29] R. A. Grusin, Premediation, Criticism 46 (2004) 17 – 39.

[30] I. Spada, V. Patti, M. Lai, IntersHate: un corpus italiano per lo studio di misoginia e intersezionalità in Twitter, Technical Report, Department of Computer Science, University of Turin, Italy, 2020. URL: https://github.com/ivsnp/inters8/blob/main/spada_thesis_IntersHate.pdf.

A. Online Resources

The annotated corpus and the guidelines are available on GitHub at the following link: https://github.com/ivsnp/inters8.