QUEEREOTYPES: A Multi-Source Italian Corpus of Stereotypes towards LGBTQIA+ Community Members 🏳️‍🌈

Alessandra Teresa Cignarella[♡,♣], **Manuela Sanguinetti**[♢], **Simona Frenda**[♡,♣], **Andrea Marra**[♡], **Cristina Bosco**[♡] **and Valerio Basile**[♡]

♡ Computer Science Department, University of Turin, Italy
♣ aequa-tech, Turin, Italy
♢ Department of Mathematics and Computer Science, University of Cagliari, Italy
{alessandrateresa.cignarella | simona.frenda | andrea.marra}@unito.it
{cristina.bosco | valerio.basile}@unito.it
manuela.sanguinetti@unica.it

Abstract

The paper describes a dataset composed of two sub-corpora from two different sources in Italian. The QUEEREOTYPES corpus includes social media texts regarding LGBTQIA+ individuals, behaviors, ideology and events. The texts were collected from Facebook and Twitter in 2018 and were annotated for the presence of stereotypes, and orthogonal dimensions (such as hate speech, aggressiveness, offensiveness, and irony in one sub-corpus, and stance in the other). The resource was developed by Natural Language Processing researchers together with activists from an Italian LGBTQIA+ not-for-profit organization. The creation of the dataset allows the NLP community to study stereotypes against marginalized groups, individuals and, ultimately, to develop proper tools and measures to reduce the online spread of such stereotypes. A test for the robustness of the language resource has been performed by means of 5-fold cross-validation experiments. Finally, text classification experiments have been carried out with a fine-tuned version of AlBERTo (a BERT-based model pre-trained on Italian tweets) and mBERT, obtaining good results on the task of stereotype detection, suggesting that stereotypes towards different targets might share common traits.

Keywords: LGBTQIA+, Stereotypes, Corpus, Italian

1. Introduction

According to a recent survey from Amnesty International USA, Twitter (now "X") is failing to protect LGBTQIA+ organizations and individuals that advocate for members of the queer community from online violence and abuse.[1] Indeed, the hatred on microblogging platforms and social media towards vulnerable communities is widespread; consequently, many projects and campaigns address this phenomenon at large. While some focus on the more general concept of Hate Speech (HS), others are more specific and address hateful phenomena, generally classifying it on the basis of its target – referring to the so-called "protected classes" to indicate especially discriminated social groups. We can thus distinguish projects that study racism, sexism, misogyny, islamophobia, and homophobia. The latter is the focus of interest for the purpose of this work.

The European Union and the Council of Europe have devoted resources to confront discrimination and hostility towards people on the basis of their sexual orientation, sexual identity or gender identity.[2] A few examples of the international scenario include the No Hate Speech Movement[3] promoted by the Council of Europe, launched in 2013 and still active with several campaigns; the BRICkS project[4] supported by the European Union; the "Silence Hate" campaign[5] launched by Amnesty International. Concerning the Italian context, it is worth mentioning the "Jo Cox" Parliamentary Commission on Intolerance, Xenophobia, Racism and Hate, active in 2016 and 2017[6]. Furthermore, on the Italian ground, in 2020, a law proposal that criminalizes discrimination, violence, and incitement to hatred when it is motivated by sexual orientation, sexual identity or gender identity, as well as by ableism has been presented and rejected by the Parliament after a public debate lasting two years.[7]

[1] https://www.amnestyusa.org/press-releases/hateful-and-abusive-speech-towards-lgbtq-community-surging-on-twitter-under-elon-musk/

[2] https://www.coe.int/en/web/human-rights-channel/LGBTQI-human-rights-Council-of-Europe

[3] https://www.coe.int/en/web/no-hate-campaign

[4] https://www.bricks-project.eu/

[5] https://www.amnesty.it/pubblicazioni/silence-hate-media-education-e-hate-speech-quaderno-di-lavoro/

[6] https://www.camera.it/leg17/1264

[7] https://www.senato.it/service/PDF/PD

LREC-COLING 2024, pages 13429–13441

Therefore, nowadays, there is no law in Italy that protects people from these forms of discrimination.

The aforementioned initiatives often involve actors from different fields, including human rights workers and representatives of local and national institutions. Taking into account the complexity of this task, we aim to tackle a research gap of abusive language detection for the Italian language, especially focused on its deeply rooted connections with **stereotypes**.

As for Hate Speech, several resources exist covering a wide range of languages, topics, and approaches (Poletto et al., 2020; Vidgen and Derczynski, 2020): most of them were created and made available in the last ten years, emphasizing how topical this subject is and how quickly is being tackled by the scientific community. English, as could be expected, is the most studied language, but many other languages are covered as well (Ross et al., 2017; Fišer et al., 2017; Sanguinetti et al., 2018; Steinberger et al., 2017), (Del Vigna et al., 2017; Corazza et al., 2019; Akhtar et al., 2019). Despite the broad interest recently raised in hateful or abusive language within the NLP community, homophobic language and online verbal abuse, precisely based on **gender identity and sexual orientation**, are starting to receive a specific focus only in very recent times.

2. Related Work

To the best of our knowledge, there are very few research attempts and resources, within the NLP community, in which homophobia has been investigated in-depth as a phenomenon on its own. Among them, it is worth mentioning the resource published by Ljubešić et al. (2019), consisting of a dataset of Facebook comments covering migrants and LGBTQIA+ and manually labeled with regard to several types of socially unacceptable discourse (Slovene and English). Carvalho et al. (2022) present a new Twitter dataset created to analyze online hate towards the most representative minorities in Portugal, namely the African descent and the Roma communities, and the LGBTQ+ community (Portuguese).

The most recent works in the NLP research community that have been dedicated to addressing homophobia and transphobia are: the 2022 Shared Task on Homophobia/Transphobia Detection in social media comments, which was conducted at the *Language Technology For Equality Diversity Inclusion workshop*, focusing on YouTube comments in English and Tamil (Chakravarthi et al., 2022); a significant study by Locatelli et al. (2023) undertook a cross-lingual study of homo-transphobia on Twitter, introducing a comprehensive taxonomy to categorize and understand public discourse surrounding LGBTQIA+ topics; and, finally, a work by Nozza et al. (2022) who delved into the negative impact of Large Language Model Models on LGBTQIA+ individuals, thus emphasizing the critical need for ethical considerations in the deployment of NLP techniques.

Another relevant activity, especially for the language studied in the present work (Italian), is the HODI shared task dedicated to the detection of hateful content towards LGBTQIA+ community members in tweets (Nozza et al., 2023).

The authors of this paper, also acknowledge the existence of TWEER, a corpus of 5,660 Italian tweets annotated according to the scheme presented in Sanguinetti et al. (2018).[8] However, the resource has not been published nor made available at the time of writing.

3. Paper contribution

With this work, we aim at filling a gap in the current state of the art by providing the following contributions:

1. **a multi-source dataset for the study of stereotypes** towards LGBTQIA+ community members in Italian, focusing on the Italian language (see Section 4)
2. **a set of classification experiments** performed by cross-validating and fine-tuning two different *pre-trained language models* (PLMs) compared to the baseline measures of the HaSpeeDe2 shared task (Sanguinetti et al., 2020) in which we show how stereotypes towards different targets share common traits, therefore, the same models could be employed for detection of stereotypes towards other vulnerable groups (see Section 5)

We believe that this work may contribute to the development of tools to counter abusive phenomena online, and help conduct research for socially good causes, with the aim of providing 'real-life' help towards vulnerable communities.

A Data Statement finally complements the description of the language resource, according to the Version 2 Scheme proposed by McMillan-Major et al. (2023) (see here: https://t.ly/Jdn5I).

4. Dataset Construction

In this section, we describe the procedure used in collecting the data, including full descriptions

FServer/DF/356433.pdf

[8] https://www.linguisticamente.org/hate-speech-su-twitter-misurare-lomofobia-e-possibile/.

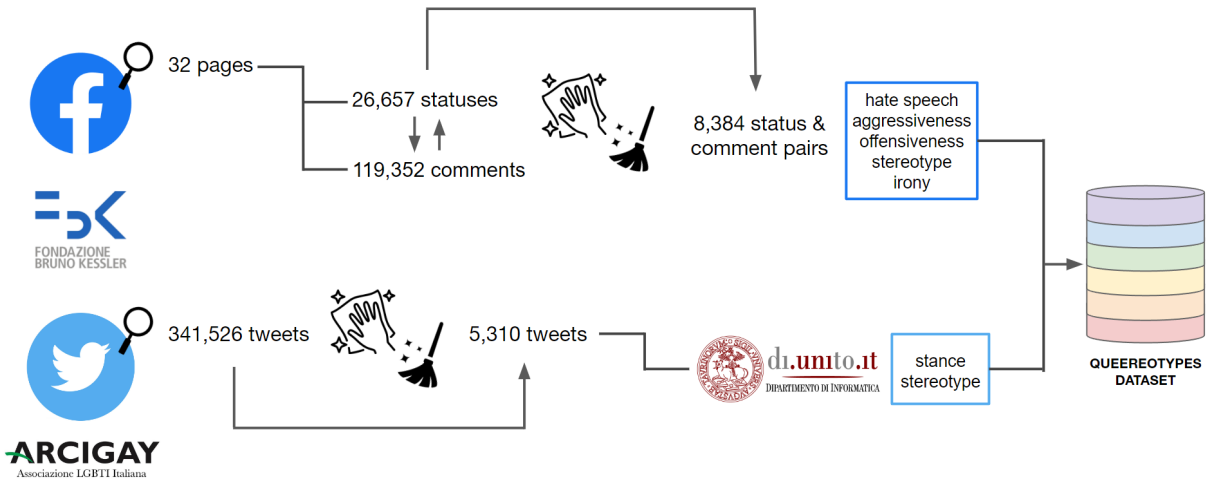

Figure 1: Pipeline of the creation of the multi-source QUEREEOTYPES dataset.

of any computational pre-processing. As the dataset described in this paper is multi-source, we will list below two different sections, one for each sub-corpus of the QUEEREOTYPES dataset: (1) QUEEREOTYPES_FB, containing posts and comments from Facebook; and (2) QUEEREOTYPES_TW, containing tweets from Twitter.

The Facebook data was originally collected by the Data Science group of *Fondazione Bruno Kessler* (FBK), and was cleaned and annotated within the scope of a Master's thesis[9] developed in the University of Turin (co-supervised by two of the authors of this manuscript) to study the phenomenon of hate speech towards LGBTQI individuals. The Twitter data was also originally collected by researchers from *FBK* within a collaboration with the *Arcigay*[10] association within the scope of the EU-funded *ACCEPT*[11] project aimed at increasing the social acceptance of LGBTQIA+ people and help Civil Society Organizations and Public Institutions to prevent homophobic and transphobic discrimination and hate in Italy.

With this contribution, we aim at joining the research efforts of two different projects that are devoted to a similar goal. First, by merging two datasets and refining them in order to be used for Natural Language Processing tasks, such as the study of hateful language and stereotypes/negative biases towards LGBTQIA+ individuals. In particular, we believe useful insights could emerge from the comparison between the two sub-corpora, which are of a same textual genre, but from different sources (Facebook and Twitter), thus encompassing multiple views and perspectives from different kind of users. This motivated the creation of this multi-source dataset, that we henceforth call QUEEREOTYPES.

4.1. QUEEREOTYPES_FB

a) Data collection and annotation scheme. For this sub-corpus, we first manually selected 32 Facebook pages overall, based on the main topics addressed and their content, both against and in support of LGBTQIA+ stances. We included pages of right-wing politicians and pro-family movements on the one hand, and groups or politicians that have been active in the defense of LGBTQIA+ rights on the other. We only used data from public pages, following consolidated collection procedures from the recent literature, e.g., (Bosco et al., 2018; Lamprinidis et al., 2021; Bosco et al., 2023).

The total number of statuses and comments originally retrieved was 26,657 and 119,352, respectively. As this part of the corpus was originally collected and annotated within the scope of a Master's thesis we inevitably had to reduce the number of texts to a tractable size for human annotators to label in a reasonable amount of time.

A random sampling was performed and the resulting dataset consists of 8,384 Facebook posts. This amount of posts belongs to the 6 pages reported in Table 1, which are also the pages that were posting the most content among the 32 originally retrieved.

The resulting dataset contains Facebook posts

[9] Franco P. (2017/2018). *Hate speech contro la categoria LGBTQI in un corpus estratto da Facebook*. Master Thesis. University of Turin.

[10] https://www.arcigay.it/

[11] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-details/31076817/777354/REC.

[12] https://www.facebook.com/gaypuntoit

[13] https://www.facebook.com/profile.php?id=100064652413413

Comments	Page Name
3,047	Gay.it[12]
1,872	Sentinelle in Piedi[13]
1,791	CitizenGo Italia[14]
846	Popolo della Famiglia[15]
698	ArciLesbica Nazionale[16]
130	Noi sosteniamo il presidente Putin contro le adozioni alle coppie gay[17]

Table 1: Number of comments extracted from each Facebook page.

and comments published between February 2018 and June 2018, and annotated according to the scheme proposed in Poletto et al. (2017) and Sanguinetti et al. (2018), considering the following dimensions:

- hate speech (yes / no)
- stereotype (yes / no)
- aggressiveness (absent / weak / strong)
- offensiveness (absent / weak / strong)
- irony (yes / no)

Following, an example from the FB sub-corpus:

• **Status message**: "God created Woman so that we all had a Mother" 🌹
• **Comment message**: the woman also has a brain, and she does whatever the f**k she wants with it.
–
Hate Speech: no,
Stereotype: no,
Aggressiveness: weak,
Offensiveness: absent,
Irony: no.

b) Annotation and inter-annotator agreement. Firstly, a sample of 60 Facebook posts has been annotated by three linguists (among the authors of the paper), in order to test the validity and the mutual comprehension of the guidelines. In Table 2 we report the values of Fleiss' κ, used to evaluate the results of this step of the annotation campaign. As it can be observed, the IAA for Hate Speech is quite high (0.7252), while the other four annotated dimensions show a lower agreement. Although, κ's scores are over 0.400, calculated in the domain of highly-subjective tasks such as Aggressiveness, Offensiveness, Irony, and Stereotype can still be considered promising.

After this first pilot annotation, a good consensus on common guidelines was reached among annotators: disagreeing cases were thoroughly discussed and some clarifications were introduced in the guidelines); then two skilled linguists conducted the annotation on the entire Facebook sub-corpus (8,384 posts). In Table 3 we report the values of Cohen's κ obtained between the two annotators.

Dimension	Fleiss' κ
Hate Speech	0.7252
Aggressiveness	0.3989
Offensiveness	0.4881
Irony	0.4177
Stereotype	0.4396

Table 2: IAA calculated with Fleiss' κ on a first sample of 60 posts from the Facebook sub-corpus.

Dimension	Cohen's κ
Hate Speech	0.6309
Aggressiveness	0.5522
Offensiveness	0.6517
Irony	0.6641
Stereotype	0.6603

Table 3: IAA calculated with Cohen's κ on the entire Facebook sub-corpus (8,384 posts).

From Table 3, it can be observed how the κ lowered for the dimension of Hate Speech (with respect to Table 2), but raised for all other dimensions, with $\kappa > 0.55$ across all annotated phenomena – which is a substantially good agreement especially considering how long and lexically complex Facebook posts can be, compared to other shorter textual genres (such as tweets). The remaining cases of disagreement were solved by a third skilled annotator.

c) Label distribution. Finally, the resulting dataset, considered as 'gold standard' among all five dimensions, consists of 2,888 Facebook posts. In Table 4 we show the label distribution across the five dimensions. The table has been split to more easily show the label distribution of the binary dimensions (Hate Speech, Stereotype, and Irony) on one side, and those featuring a three-way scheme on the other (Aggressiveness and Offensiveness).

It can be seen how the *negative* classes (i.e., the classes where the annotated phenomenon is not present), are always the majority. For instance, there is only 7.8% of Hate Speech, only 5.3% of Stereotypes, 14.1% of Irony. Slightly higher values are shown for Aggressiveness and Offensive-

[14] https://www.facebook.com/CitizenGOItalia/
[15] https://www.facebook.com/popolodellafamiglia
[16] https://www.facebook.com/Arcilesbica
[17] Page deleted at the time of writing (April/June 2023).

	Hate Speech		Stereotype		Irony	
yes	224	7.8%	152	5.3%	408	14.1%
no	2,664	92.2%	2,736	94.7%	2,480	85.9%
Total			**2,888**			

	Aggressiveness		Offensiveness	
strong	100	3.5%	109	3.7%
weak	634	22.0%	408	14.1%
absent	2,154	74.6%	2,371	82.1%
Total		**2,888**		

Table 4: Label distribution in the Facebook sub-corpus.

ness (25.5% and 17.8% summing weak and strong cases). The low percentages are not surprising, as they are broadly in line with other studies on Italian social media data that collected similar kinds of data by using non-polarized keywords (Sanguinetti et al., 2020; Bosco et al., 2023).

4.2. QUEEREOTYPES_TW

a) Data collection and annotation scheme. This sub-corpus contains tweets collected by activists from an Italian LGBTQIA+ association using the Search Twitter API v1.1[18] from March 2018 to November 2018. Data have been retrieved by using keywords such as *gay*, *lesbian*, *pride*, *gender*; the resulting collection included 341,526 tweets overall (see Figure 1). After filtering, cleaning and resizing for a reasonable amount of data to be manually annotated, the final size of the Twitter sub-corpus consisted of 5,310 tweets in Italian.

Four NLP researchers (among the authors of this paper) enriched this data collection with two layers of annotation, marking the presence of stereotypical expressions on one side and that of stance on the other.

The initial layer of annotation is crucial for examining the dimension of **stereotypes** concerning LGBTQIA+ individuals and establishing a link with the Facebook sub-corpus of the QUEEREOTYPES dataset. In contrast, the annotation of stance is driven by its significance: we aim to analyze the positions conveyed in the conversations' messages concerning individuals, behaviors, ideologies, and events related to the LGBTQIA+ community.

For the annotation of stereotypes, we adhered to the guidelines used for annotating this aspect in the Facebook sub-corpus, as described in Section 4.1. For the stance dimension, we employed a 5-point scalar annotation from -2 to +2 (Küçük and Can, 2020), chosen for its capacity to capture the granularity of the phenomenon and its potential for seamless aggregation into a 3-label format, as showed by Mohammad et al. (2016), if necessary. To summarize, the Twitter sub-corpus consists of the two following dimensions:

- stereotype (yes / no)
- stance (-2 / -1 / 0 / +1 / +2)

Following, an example from the TW sub-corpus:

• **Tweet text:** ...or they invent the "LGBT" party to pretend they want to protect their rights to devastate everyone's minds and bodies with gender ideology.
–
Stance: against,
Stereotype: yes.

b) Annotation and inter-annotator agreement. The agreement regarding stance has been calculated using a weighted version of Krippendorff's α, where the difference between, e.g., -2 and -1 is considered less impactful than the difference between -2 and +1 (Antoine et al., 2014). The agreement on stereotypes has been calculated with the base version of Krippendorff's α. In Table 5 we display the values of the IAA obtained among the 4 expert linguists, regarding both dimensions.

Dimension	Krippendorff's α
Stereotype	0.5073
Stance	0.4783

Table 5: IAA calculated with Krippendorff's α on the Twitter sub-corpus.

Although stance was annotated with a five-point scale (-2 / -1 / 0 / +1 / +2) to guarantee a fine-grained granularity of phenomenon, for an easier computation of the label distribution and in light of the experiments described in Section 5, these labels were finally mapped to *against*, *neutral* and *favor* following Mohammad et al. (2016); Küçük and Can (2020).

c) Label distribution. The resulting dataset consists of 3,427 tweets. In Table 6, we show the label distribution of both annotated phenomena. On the left is displayed the value for stereotype, annotated with a binary label (*yes* / *no*), and on the right is displayed the dimension of stance, which is annotated with three labels: *against* / *neutral* / *favor*.

Also in this sub-corpus extracted from Twitter, the *negative* classes (meaning the opposite class of the phenomena we aimed at studying in this research) are the majority of the data. For instance, there is only 7.2% of Stereotypes, and only 9.0% of Stance=*against*. Slightly higher is also the value

[18] https://developer.twitter.com/en/products/twitter-api

Stereotype		
yes	248	7.2%
no	3,179	92.8%
Total	**3,427**	

Stance		
against	307	9.0%
neutral	1,613	47.1%
favor	1,507	43.9%
Total	**3,427**	

Table 6: Label distribution in the Twitter sub-corpus.

for Stance=*neutral*, i.e., 47.1%. If we analyze the dimension of Stereotype, which is annotated in both sub-corpora of the QUEEREOTYPES dataset, we will notice that in both cases the percentage showing its presence is around 5% and 7%.

4.3. QUEEREOTYPES: The Multi-Source Corpus

In order to account for sufficient textual diversity and delete duplicates, we calculated text similarity through Jaccard's coefficient[19] starting from a value of J = 0.90 and data was manually inspected. After each check, the value of J was reduced by 0.05 and followed by a new manual inspection. Finally, J = 0.55 is the threshold in which the inspection revealed a good trade-off between quality and quantity of data.

We also discarded tweets and Facebook posts that only received less than two annotations. Figure 1 summarizes the dataset collection and cleaning pipeline, from the text retrieval to the annotation of different dimensions and the merging of the two sub-corpora into the final QUEREEOTYPES dataset.

	Total number of texts	
Sub-corpus	**'Non-aggregated'**	**'Gold standard'**
Facebook	8,384	2,888
Twitter	5,310	3,427
Total	**13,694**	**6,215**

Table 7: Total number of texts with 'non-aggregated' labels and of texts of the 'gold standard'.

The resulting size of the "non-aggregated" and the "gold standard" multi-source annotated dataset is shown in Table 7. It can be seen how the total of texts (both Facebook posts and tweets) ever annotated is 13,694, but the consensus of the majority voting, for creating the 'gold standard' led to a smaller number: 6,215.

Considering the growing interest in the NLP community in leveraging the richness of information contained in non-aggregated labeled data, we created two different versions in which the labels applied by each individual annotator are also available. In line with the perspectivist data manifesto[20], this characteristic of our data will be able to model unique points of view on the studied phenomenon. This double facet of the dataset is a richness considering the growing branch of research in which NLP models are not solely trained on a gold standard, but rather on disagreeing annotations: i.e, *learning with disagreement* (Uma et al., 2021; Akhtar et al., 2019).

5. Experiments

In this section, we first present a technical validation of the multi-source QUEEREOTYPES dataset, and secondly a binary classification task for the detection of stereotypes.

5.1. Technical Cross-Validation

The dataset has been validated in terms of its robustness as a training set for supervised predictive models for NLP tasks. For the Facebook sub-corpus, we performed five classification experiments, one for each of the annotated dimensions. Each experiment consists of a 5-fold cross-validation, and the results are presented in terms of precision, recall and F1-score (i.e., the harmonic mean of precision and recall) for each of the classes. Additionally, the *macro-averaged* precision (P), recall (R) and F1-score (F1) are shown, as a summary of the performance of the model on each dimension, also for better comparison with the state of the art in these NLP tasks. The model is a fine-tuned version of AlBERTo (Polignano et al., 2019), a BERT model (Devlin et al., 2019) pre-trained on a large collection of tweets in the Italian language (Basile et al., 2018), with a fully-connected layer for the output, with `softmax` activation, a learning rate of 10^{-6}, and a batch size of 32. For each fold, the model is trained for 5 epochs. Moreover, the experiment was repeated five times and the results were averaged, to alleviate the differences due to the random initialization of the neural network.

Tables 8–12 show the results of this evaluation. The performance varies depending on the observed phenomenon. In general, class imbalance is one of the main issues, with the less represented classes (such as the positive class in the Hate Speech classification) being severely penalized especially in terms of precision.

For the Twitter sub-corpus, the dimension of Stance is annotated as a scalar value from -2 to +2, rather than a label. As such, the appropriate experiment would be a regression evaluated by

[19] https://www.kaggle.com/code/jfaucett/nlp-tutorial-0001-exploring-jaccard-similarity

[20] https://pdai.info/

Class	P	R	F1
no	.949	.852	.895
yes	.220	.443	.278
macro-average	.585	.648	.587

Table 8: Results of the 5-fold cross-validation on the QUEEREOTYPES_FB sub-corpus: Hate Speech.

Class	P	R	F1
absent	.871	.702	.768
weak	.372	.491	.408
strong	.164	.359	.193
macro-average	.469	.517	.457

Table 9: Results of the 5-fold cross-validation on the QUEEREOTYPES_FB sub-corpus: Aggressiveness.

Class	P	R	F1
absent	.919	.815	.861
weak	.348	.456	.377
strong	.296	.468	.340
macro-average	.521	.579	.526

Table 10: Results of the 5-fold cross-validation on the QUEEREOTYPES_FB sub-corpus: Offensiveness.

Class	P	R	F1
no	.972	.829	.893
yes	.160	.561	.244
macro-average	.566	.695	.569

Table 11: Results of the 5-fold cross-validation on the QUEEREOTYPES_FB sub-corpus: Stereotype.

Class	P	R	F1
no	.922	.771	.835
yes	.320	.596	.402
macro-average	.621	.683	.619

Table 12: Results of the 5-fold cross-validation on the QUEEREOTYPES_FB sub-corpus: Irony.

means of a correlation metric. However, in an effort to provide a more interpretable result, we transform this task into a three-way classification task, where the negative scores are mapped to the label *against*, the positive scores are mapped to the label *favor*, and the zero is mapped to the label *neutral*, as previously described in Section 4. The stereotype dimension is already categorical. The experimental setting is the same as for the Facebook sub-corpus, including the same model and hyperparameters described above.

Tables 13 and 14 show the results of the cross-validation experiment on the Twitter sub-corpus. We can see how the class imbalance is an issue in this sub-corpus as well, although the performance,

Class	P	R	F1
against	.500	.408	.450
favor	.670	.668	.669
neutral	.686	.716	.701
macro-average	.619	.597	.606

Table 13: Results of the 5-fold cross-validation on the QUEEREOTYPES_TW sub-corpus: Stance.

Class	P	R	F1
no	.941	.951	.946
yes	.279	.240	.258
macro-average	.610	.596	.602

Table 14: Results of the 5-fold cross-validation on the QUEEREOTYPES_TW sub-corpus: Stereotype.

for both classification tasks, seems promising.

5.2. Stereotype Classification

Finally, we carried out a number of experiments on the automatic detection of stereotypes. Considered that during HaSpeeDe2 at EVALITA 2020, the organizers proposed a binary classification task for the detection of stereotypes in Italian tweets (Sanguinetti et al., 2020), we use the dataset provided in the competition to obtain a baseline measure, against which we compare our proposed method. We frame two different experimental settings:

1. *HaSpeeDe2 Setting*, in which we train and test models on the training and test set provided within the HaSpeeDe2 shared task for the binary classification of stereotype (Sanguinetti et al., 2020);

2. *Expanded Setting*, in which we add QUEEREOTYPES to the training dataset of HaSpeeDe2 for training a model, and we test it against the same test set provided from the evaluation campaign.

Table 15 displays the averaged results of the two models on 5 runs, mBERT (Devlin et al., 2019) and AlBERTo (Polignano et al., 2019) evaluated in both settings described above.

	HaSpeeDe2 Setting			*Expanded Setting*		
	P	**R**	**F1**	**P**	**R**	**F1**
mBERT	.740	.719	.698	.739	.740	.735
AlBERTo	.751	.729	.716	.746	.744	**.744**

Table 15: Results of textual classification experiments on the Stereotype dimension with mBERT and AlbERTo.

Furthermore, we computed two baseline measures. First, a majority class baseline MCB (F1

= 0.355) and secondly, a random baseline RB (F1 = 0.504). Moreover, it is noteworthy that the only currently existing benchmark for stereotype detection in Italian is derived from subtask B of HaSpeeDe2 (Sanguinetti et al., 2020), in which a baseline SVC is computed, obtaining F1 = 0.715, and a second Baseline MFC is calculated with F1 = 0.355.[21] Please, note that in HaSpeeDe2 stereotypes are expressed towards a different target (i.e., migrants).

Looking back at Table 15, both models show similar precision and recall values, with AIBERTo slightly outperforming mBERT in terms of precision, recall, and F1-score, in both settings. The outcome is supported by the fact that AIBERTo is pre-trained on a large collection of texts belonging to the same textual genre as the dataset used for classification experiments (i.e., tweets), and on Italian solely. On the other hand, mBERT is multilingual and pre-trained on a bigger variety of textual genres.

In the *Expanded Setting*, the results show that models fine-tuned on a broader training set improve their performance, with Recall and F1-score being higher than in the *HaSpeeDe2 Setting*. This suggests that incorporating the additional QUEEREOTYPES data enhances the models' tendency to identify more instances as positive examples of Stereotype.

Considering "The addition of the QUEEREOTYPES dataset is not beneficial for the performance of the model" as a null hypothesis (H0), we conducted significance tests with the two different models employed in our experimental setting. For the mBERT model, we calculated a p-value of 0.0372, and for the AIBERTo model, we obtained a p-value of 0.0140. In both cases, the resulting p-values are less than the chosen significance level of 0.05. Consequently, we can reject the null hypothesis (H0) for both models and conclude that the addition of the QUEEREOTYPES data has benefited the performance of both mBERT and AIBERTo models.

It is worth stressing that the HaSpeeDe2 dataset encodes the presence of stereotypes in Italian tweets towards immigrants, Muslims and Roma (Sanguinetti et al., 2020), while, on the other hand, QUEEREOTYPES encodes stereotypes towards LGBTQIA+ individuals. Overall, the obtained results highlight the importance of dataset diversity and extension in training models to enhance performance. The higher scores obtained in the *Expanded Setting*, where both models were fine-tuned on both racist and homophobic stereotypes, seem also to point out that stereotypes towards different targets share common traits, therefore, the phenomenon of 'stereotyping' could be more generalizable, and the same models might be employed also for detection of stereotypes towards other vulnerable groups (women, elderly, disabled bodies, non-white people, ethnic minorities, homeless, etc.).

[21] For the results of HaSpeeDe2 in detail, please refer to Table 7 in this paper: https://pure.rug.nl/ws/portalfiles/portal/155321945/paper162.pdf.

6. Conclusions

We have presented the first Italian multi-source dataset that is devoted to the study of stereotypes against LGBTQIA+ individuals, and in which a collaboration with members of such community is established for data collection. It is a precious resource to be used in NLP experiments as well as for linguistic and lexical studies. The QUEEREOTYPES dataset includes 13,694 social media texts regarding LGBTQIA+ members, behaviors, ideology, and events collected from both Facebook and Twitter. The data have been collected and organized according to the *perspectivist data manifesto*, thus preserving the non-aggregated labels from every single annotator.

The dataset underwent comprehensive validation through 5-fold cross-validation experiments, across all dimensions, ensuring models' ability to generalize effectively to unseen data, fostering robustness in their predictions. Furthermore, we carried out a set of computational experiments, on the automatic detection of stereotypes towards LGBTQIA+ people, by employing a fine-tuned version of AIBERTo and mBERT. The results show that incorporating the additional QUEEREOTYPES data enhances the models' tendency to identify more instances as positive examples of stereotype. Furthermore, the obtained results highlight the importance of dataset diversity and extension in training models to improve performance.

Finally, the outcome of our study, seem also to point out that stereotypes towards different targets share common traits, therefore, the same models might also be employed for detection of stereotypes towards other vulnerable groups. Besides allowing the study of stereotypes in NLP, this resource might be helpful to prevent discrimination towards marginalized communities.

Data and Code Availability

The QUEEREOTYPES dataset (v1) is accessible at a Open Science Framework repository at this link: <QUEEREOTYPES - Open Science Framework>, to researchers on a "sharing-with-peers" basis. Interested parties will be required to reach out to the authors, complete a form, and sign an agreement contract, outlining the specifics of their research in

order to obtain the password that protects the files. It is essential for them to ensure compliance with GDPR regulations and other policies from both X (former Twitter) and Facebook, as well as following open-source practices to ensure the possibility of data-driven NLP studies in the spirit of the Language and Resources Evaluation Conference.

Limitations

The dataset used in this study was collected during 2018. Since online discourse and attitudes can greatly vary over time, the findings and conclusions drawn from this dataset may not reflect the current landscape of stereotypes and online behavior towards LGBTQIA+ individuals, since many relevant events occurred ever since (one among many: the COVID-19 pandemic).

The dataset focuses specifically on Italian texts, limiting its generalizability to other languages and cultures. The stereotypes, expressions, and contextual nuances present in Italian social media may not align with those found in different linguistic and cultural contexts. Thus, caution should be exercised when extrapolating the findings to other languages or cultures.

The paper reports the use of a fine-tuned version of the AlBERTo and mBERT models for text classification experiments. The performance and results obtained may be influenced by the specific characteristics of these models and their training data. Different models or approaches might yield different results, and the generalizability of the findings to other models or architectures should be further investigated.

Although the dataset development involved collaboration with an Italian LGBTQIA+ non-profit organization, the extent and nature of the involvement may vary. The limitations or biases arising from the dataset creation process, including data collection and annotation, should be considered in light of the specific involvement of the activist group and potential power dynamics that might have influenced the dataset's construction.

Ethical Considerations

The study presented in the paper raises several important ethical considerations that should be carefully addressed in the collection, analysis, and dissemination of the dataset and findings.

This study and the creation of the QUEEREOTYPES dataset aim to analyze stereotypes towards LGBTQIA+ individuals. However, in collecting and annotating the dataset, there is a risk of reinforcing or perpetuating existing biases and stereotypes. Researchers must be vigilant in their approach to avoid amplifying harmful narratives or stigmatizing the LGBTQIA+ community further. Careful consideration should be given to the potential impact of the research on marginalized communities and the broader social implications of reinforcing stereotypes.

Indeed, creating such a dataset aims to develop tools and measures to reduce the online spread of stereotypes. While this is a laudable goal, it is important to consider the potential misuse or unintended consequences of such tools. Care should be taken to avoid deploying systems that may inadvertently censor legitimate speech or disproportionately target certain individuals or communities. A thorough analysis of the ethical implications of the developed tools should be conducted to minimize harm and ensure fairness.

To ensure responsible and ethical usage, we intend to implement mechanisms to track the utilization of the dataset. By keeping a record of who accesses and uses the dataset, we aim to promote a better understanding of its impact, foster collaboration, and potentially address any concerns that may arise from its usage, and it will be made available exclusively for research purposes.

Acknowledgements

The authors would like to thank Paola Franco because this works stems directly from her Master's Thesis discussed in the University of Turin, and Fabio Poletto for his engagement in the early stages of this research and for his contribution with annotations. Furthermore, the authors want to extend their thanks to Marco Cristoforetti and Cesare Furlanello from *Fondazione Bruno Kessler* and all the other people who were involved in the ACCEPT project.[22] Finally, many thanks also to Shamar Droghetti, *ARCIGAY* and to the activists who personally took part in this collaborative research.

The work of Alessandra Teresa Cignarella and Cristina Bosco was partially supported by the International project 'STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES' funded by the Compagnia di San Paolo and Volkswagen Stiftung under the 'Challenges for Europe' call for Project (CUP: B99C20000640007). The work of Cristina Bosco is also partially funded by Compagnia di San Paolo - Bando ex-post 2020 - SteroetypHate. The work of Valerio Basile is partially supported by Compagnia di San Paolo - Bando ex-post 2020 - "Toxic Language Understanding in

[22] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-details/31076817/777354/REC.

Online Communication - BREAKhateDOWN". The work of Simona Frenda was partially funded by the 'Multilingual Perspective-Aware NLU' project in partnership with Amazon Alexa. The work of Manuela Sanguinetti was partially supported by the project DEMON "Detect and Evaluate Manipulation of ONline information" funded by MIUR under the PRIN 2022 grant 2022 BAXSPY (CUP F53D23004270006, NextGenerationEU).

Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A New Measure of Polarization in the Annotation of Hate Speech. In *Proceedings of the International Conference of the Italian Association for Artificial Intelligence*, pages 588–603.

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 550–559, Gothenburg, Sweden. Association for Computational Linguistics.

Dennis Assenmacher, Indira Sen, Leon Fröhling, and Claudia Wagnero. 2023. The end of the rehydration era - the problem of sharing harmful twitter research data. In *2nd Workshop on Novel Evaluation Approaches for Text Classification Systems (NEATCLasS)*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018.*

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Cristina Bosco, Felice dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the evalita 2018 hate speech detection task. In *Proceedings of EVALITA '18, Evaluation of NLP and Speech Tools for Italian*, Turin, Italy.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Erik Cambria, Marco Grassi, Amir Hussain, and Catherine Havasi. 2012. Sentic computing for social media marketing. *Multimedia tools and applications*, 59(2):557–577.

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *IberEval@SEPLN*. CEUR-WS.org.

Paula Carvalho, Bernardo Cunha, Raquel Santos, Fernando Batista, and Ricardo Ribeiro. 2022. Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2362–2370.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Philip McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. Cross-platform evaluation for Italian hate speech detection. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International Conference on Web and Social Media*, pages 512–515. AAAI.

Emiliana De Blasio, Donatella Selva, Michele Sorice, et al. 2022. Il dibattito sul DDL Zan e la post-sfera pubblica italiana. *Mediascapes journal*, 19(1):89–112.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech

Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20. Association for Computational Linguistics (ACL).

Rogers Prates de Pelle and Viviane P Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95. CEUR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Johan Fernquist, Oskar Lindholm, Lisa Kaati, and Nazar Akrami. 2019a. A study on the feasibility to detect hate speech in swedish. In *2019 IEEE international conference on big data (Big Data)*, pages 4724–4729. IEEE.

Johan Fernquist, Oskar Lindholm, Lisa Kaati, and Nazar A. Akrami. 2019b. study on the feasibility to detect hate speech in swedish. In *2019 IEEE international conference on big data (Big Data)*, pages 4724–4729. IEEE.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

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)*. CEUR-WS.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN*. CEUR-WS.org.

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

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics.

Davide Locatelli, Greta Damo, and Debora Nozza. 2023. A Cross-Lingual Study of Homotransphobia on Twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. Data statements: From technical concept to community practice. *ACM J. Responsib. Comput.* Just Accepted.

JA Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017a. Abusive language detection on

arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017b. Abusive language detection on arabic social media. *Proceedings of the first workshop on abusive language online*, pages 52–56.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*, pages 145–153. International World Wide Web Conferences Steering Committee.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, Dirk Hovy, et al. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, , Caselli Tommaso, and Viviana Patti. 2023. HODI at EVALITA 2023: Overview of the EVALITA 2023 task on Homotransphobia Detection in Italian Tweets. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, page (in press). CEUR-WS.org.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018a. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 1.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018b. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 221–230. AAAI.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an Italian Twitter corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, volume 2006, Rome, Italy. CEUR Workshop Proceedings.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Francisco Rangel and Paolo Rosso. 2019. On the implications of the general data protection regulation on the organisation of evaluation tasks. *Language and Law/Linguagem e Direito*, 5(2):95–117.

Manuela Sanguinetti, Gloria Comandini, Elisa di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, Irene Russo, and Pisa. 2020. HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR-WS.org.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS One*, 15.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142. Association for Computational Linguistics (ACL).

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017. *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

Language Resource References

Bosco, Cristina and Felice, Dell'Orletta and Poletto, Fabio and Sanguinetti, Manuela and Maurizio, Tesconi and others. 2018. *Overview of the evalita 2018 hate speech detection task*. CEUR.

Bosco, Cristina and Patti, Viviana and Frenda, Simona and Cignarella, Alessandra Teresa and Paciello, Marinella and D'Errico, Francesca. 2023. *Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP*. Elsevier.

Carvalho, Paula and Cunha, Bernardo and Santos, Raquel and Batista, Fernando and Ribeiro, Ricardo. 2022. *Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal*. PID https://aclanthology.org/2022.lrec-1.253.pdf.

Fišer, Darja and Erjavec, Tomaž and Ljubešić, Nikola. 2017. *Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene*. Association for Computational Linguistics (ACL). PID 10.18653/v1/W17-3007.

Lamprinidis, Sotiris and Bianchi, Federico and Hardt, Daniel and Hovy, Dirk. 2021. *Universal Joy A Data Set and Results for Classifying Emotions Across Languages*. Association for Computational Linguistics.

Ljubešić, Nikola and Fišer, Darja and Erjavec, Tomaž. 2019. *The FRENK datasets of socially unacceptable discourse in Slovene and English*. Springer.

Nozza, Debora and Cignarella, Alessandra Teresa and Damo, Greta and Caselli, Tommaso and Patti, Viviana. 2023. *HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian*.

Ross, Björn and Rist, Michael and Carbonell, Guillermo and Cabrera, Benjamin and Kurowsky, Nils and Wojatzki, Michael. 2017. *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis*. Asian Federation of Natural Language Processing. PID https://github.com/UCSM-DUE/IWG_hatespeech_public.

Sanguinetti, Manuela and Comandini, Gloria and di Nuovo, Elisa and Frenda, Simona and Stranisci, Marco and Bosco, Cristina and Caselli, Tommaso and Patti, Viviana and Russo, Irene and Pisa. 2020. *HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task*. CEUR-WS.org. PID https://github.com/msang/haspeede.

Manuela Sanguinetti and Fabio Poletto and Cristina Bosco and Viviana Patti and Marco Stranisci. 2018. *An Italian Twitter Corpus of Hate Speech against Immigrants*. ELRA. PID https://github.com/msang/hate-speech-corpus.

Steinberger, Josef and Brychcín, Tomáš and Hercig, Tomáš and Krejzl, Peter. 2017. *Cross-lingual Flames Detection in News Discussions*. INCOMA Ltd. PID 10.26615/978-954-452-049-6_089.