View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Institutional Research Information System University of Turin

Computational Linguistics Against Hate: Hate Speech Detection and Visualization on Social Media in the "Contro L'Odio" Project

Arthur T. E. Capozzi, Mirko Lai
Valerio Basile, Fabio Poletto
Manuela Sanguinetti
Cristina Bosco
Viviana Patti
Giancarlo Ruffo
University of Turin
name.surname@unito.it

Cataldo Musto
Marco Polignano
Giovanni Semeraro
University of Bari "Aldo Moro"
name.surname@uniba.it

Marco Stranisci
ACMOS
m.stranisci@acmos.net

Abstract

The paper describes the Web platform built within the project "Contro l'odio", for monitoring and contrasting discrimination and hate speech against immigrants in Italy. It applies a combination of computational linguistics techniques for hate speech detection and data visualization tools on data drawn from Twitter. It allows users to access a huge amount of information through interactive maps, also tuning their view, e.g., visualizing the most viral tweets and interactively reducing the inherent complexity of data. Educational courses for high school students and citizenship has been developed which are centered on the platform and focused on the deconstruction of negative stereotypes against immigrants, Roma, and religious minorities, and on the creation of positive narratives.

1 Introduction

Hate Speech (HS) is a multi-faceted phenomenon with countless nuances, a high degree of individual and cultural variation, and intersections with related concepts such as offensive language, threats, bullying and so on.
The detection of HS is a recent yet popular task that is gaining the attention of the NLP community but also that of public institutions and private companies. There are several problems connected with this delicate task: a cultural-dependent definition, a highly subjective perception, the need to remove potentially illegal contents quickly from the Web and the connected risk to unjustly remove legal content, the partly overlapping linguistic phenomena that make it hard to identify HS. English social media texts are the most studied, but other languages, sources and textual genres are investigated as well.

"Contro l'odio"[1] is a project for countering and preventing racist discrimination and HS in Italy, in particular focused against immigrants. On the one hand, the project follows and extends the research outcomes emerged from the 'Italian Hate Map project' (Musto et al., 2016), whose goal was to identify the most-at-risk areas of the Italian country, that is to say, the areas where the users more frequently publish hate speech, by exploiting semantic analysis and opinion mining techniques. On the other hand, "Contro l'odio" benefits from the availability of annotated corpora for sentiment analysis, hate speech detection and related phenomena such as aggressiveness and offensiveness, to be used for training and tuning the HS detection tools (Sanguinetti et al., 2018; Poletto et al., 2017). The project brings together the competences and active participation of civil society organizations Acmos[2] and Vox[3], and two academic research groups, respectively from the University of Bari and Turin.

This paper focuses on the technological core of the project, a Web platform that combines computational linguistics analysis with visualization techniques, in order to provide users with an inter-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1] https://controlodio.it/
[2] http://acmos.net/
[3] http://www.voxdiritti.it/

CEUR-WS.org/Vol-2481/paper14.pdf

active interface for exploring the dynamics of the discourse of hate against immigrants in Italian social media. Three typical targets of discrimination related to this topical focus are taken into account, namely Migrants, Muslims and Roma, since they exemplify discrimination based on nationality, religious beliefs and ethnicity, respectively. Since November 2018 the platform analyses daily Twitter posts and exploits temporal and geo-spatial information related to messages in order to ease the summarization of the hate detection outcome.

2 Related work

In the last few years several works contributed to the development of HS detection automatic methods, both releasing novel annotated resources, lexicons of hate words or presenting automated classifiers. Two surveys were recently published on this topic (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). For what concerns Italian a few resources have been recently developed drawn from Twitter (Sanguinetti et al., 2018; Poletto et al., 2017) and FaceBook (Del Vigna et al., 2017), where the annotation of hateful contents also extends the simple markup of HS. A multilingual lexicon of hate words has been also developed (Bassignana et al., 2018) called Hurtlex[4]. The lexicon, originally built from 1,082 Italian hate words compiled in a manual fashion by the linguist Tullio De Mauro (De Mauro, 2016), has been semi-automatically extended and translated into 53 languages. The lexical items are divided into 17 categories such as homophobic slurs, ethnic slurs, genitalia, cognitive and physical disabilities, animals and more.

Since 2016, shared tasks on the detection of HS or related phenomena (such as abusive language or misogyny) in various languages have been organized, benefiting from the developed datasets and effectively enhancing advancements in resource building and system development both. These include in particular HatEval at SemEval 2019 (Basile et al., 2019), AMI at IberEval 2018 (Fersini et al., 2018b), HaSpeeDe and AMI at EVALITA 2018 (Bosco et al., 2018; Fersini et al., 2018a).

The project "Contro l'odio" follows and extends the research outcome emerged from the 'Italian Hate Map project' (Musto et al., 2016), where a lexicon developed within the project (Lingiardi et al., 2019) has been exploited to provide a fine-grained classification of the nature of the hate speeches posted by the users on different hate targets. In "Contro l'odio" we inherited the idea of map-based visualization to show the distribution of the hate speech, but we enhance it in two main directions: a) by creating a web platform that enables a *daily monitoring* of hate speech against immigrants in Italy and its evolution over time and space; b) by adding a level of interactivity with the results of the automatic detection of hate speech, both in terms of maps and of hate words' inspection, which enabled interesting activities for countering hate in schools. Monitoring and countering HS is a shared goal with several recent projects, with different focuses w.r.t countries and territories monitored, targets of hate, granularity of the detection, visualization techniques provided to inspect the monitoring results. Let us mention the *CREEP* project[5] on monitoring cyberbullying online (Menini et al., 2019), with an impact also on the Italian territory, *HateMeter*[6], with a special focus on Anti-Muslim hatred online, the MANDOLA project[7] providing a reporting infrastructure enabling the reporting of illegal hate-related speech, and the *Geography of Hate* project[8] in the US.

3 The Contro l'odio monitoring platform

3.1 Architecture

The architecture consists of four main modules. The data collection module gathers the tweets by using the Stream Twitter API and filters them by keywords. The automatic classifier module automatically annotates the presence of HS in the filtered tweets, relying on a supervised approach. The next module stores the annotated tweet aggregating them by time and place in a database. The last module, implemented by relying on a *node.js* server, exposes the API that are requested by the front end (Figure 1).

3.2 Data Collection

We started collecting tweets from October 1st 2018 by using the Twitter's Stream API. The

[4] http://hatespeech.di.unito.it/resources.html

[5] http://creep-project.eu/

[6] http://hatemeter.eu/

[7] http://mandola-project.eu/

[8] http://www.antiatlas.net/geography-of-hate-geotagged-hateful-tweets-in-the-united-states-en/

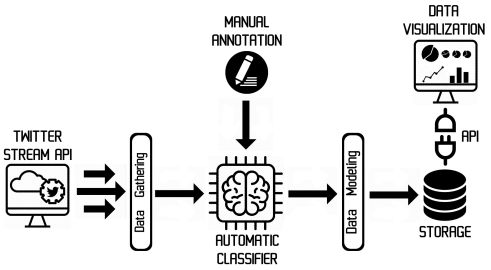

Figure 1: Architecture of the 'Contro l'odio' platform

streaming is filtered using the vowels as keywords and the alpha-2 code *it* as language filter. About $700,000$ Italian statuses are daily gathered, but only about $17,000$ are relevant for monitoring discrimination and HS against immigrants in Italy. We filtered relevant tweets by using the keywords proposed in Poletto et al. (2017), considering three typical targets of discrimination — namely migrants, Roma and religious minorities.

3.3 The Hate Detection Engine

In order to automatically label the tweets, we developed a supervised classifier to predict the presence of HS in text. The classification is binary (i.e., presence of HS vs. absence of HS). We employ a Support Vector Machine (SVM) classifier with one-hot unigram representation as feature vector. We train the classifier on the Italian Hate Speech Corpus (Sanguinetti et al., 2018, IHSC), a collection of about 6,000 tweet in the Italian language, manually annotated both by experts and crowdsourced annotators along several dimensions: hate speech, aggressiveness, offensiveness, irony, stereotype, and intensity. IHSC is particularly well suited for our scenario, since the data have been specifically collected on the topic of immigration and ethnic/religious minorities.

The following tweets are two examples of annotated tweets:

1. #dallavostraparte non ci sono moderati, sono tutti terroristi pronti a tagliarci la testa e per questo io li odio a morte!

#onyourside there are no moderates, they all are terrorists ready to cut our head off and for this I hate them to death!

2. Tanto con il sole i nomadi non vengono più a scuola. Per qualcuno questa è la soluzione...

Nomads no longer come to school when it's sunny. For some this is the solution...

In example 1, the target is "religious minorities" and the author spreads and incites violence against Islamic people (the tweet contains hate speech). In example 2, the target is "Roma", and the previous conditions are not detected, there's not hate speech. By performing cross-validation experiments on such corpus, we estimate the best hyperparameters for the model: 27,642 features, learning rate *optimal*, *linear* kernel. With this settings, we record a prediction performance in cross-validation of 0.81 (0.70 for the class *hate speech*) precision and 0.81 (0.67 for the class *hate speech*) recall ($F_{avg} = 0.80 \pm 0.01$). Recently, new classification strategies base on language understanding models have been demonstrated to be suitable for the task of hate speech detection, obtaining encouraging results. As a consequence, we are considering the possibility to compare our model with a classifier base on AlBERTo (Polignano et al., 2019) as a further step for improving the performances of our hate detection engine.

It is important to note that the Italian Hate Speech Corpus has been collected in a specific time frame, from October 1st, 2016 to April 25th, 2017 (Poletto et al., 2017). The relative distribution of topics may change over time, thus we expect a performance drop when applying the model trained on IHSC to new, recent data. In order to measure this gap, we annotated 2,000 additional tweets each month for several months, collected from the Contro l'Odio pipeline (Section 3.2) and confronted the prediction of our classifier against the manual annotation. The data have been annotated in a crowdsourcing fashion, using the online platform Figure Eight[9]. The performance of the classifier trained on IHSC on the new test set, in terms of F_{avg}, degrades as the time frame moves farther from that of of IHSC: October (0.57), November (0.56), and December (0.54), 2018, and January (0.51), and February (0.47), 2019. However, we plan to reintroduce the newly annotated datasets (this experiment is currently ongoing) in the training set and re-train the model, in order to make the system more robust across time, and to keep monitoring the performance.

4 Visualizing and Interacting with Estimated Hate

4.1 Interactive Hate Maps

The main view of the dashboard is a choropleth map and allows the user to explore the spatial

[9] https://www.figure-eight.com

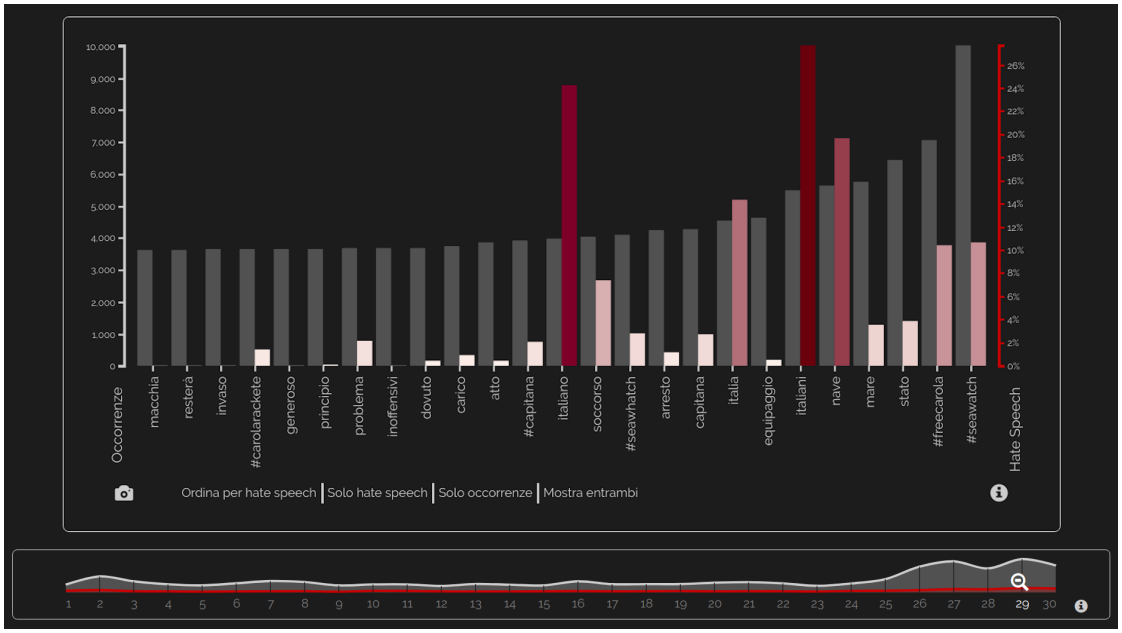

Figure 2: Word occurrences bar chart.

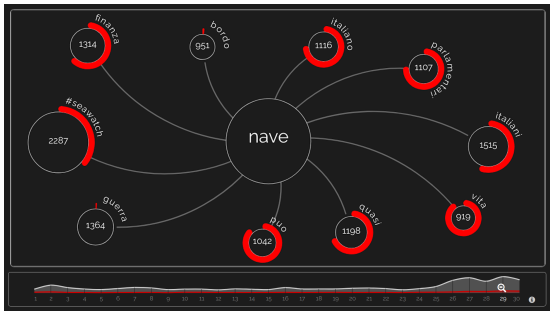

Figure 3: Word co-occurrences network.

dimension (regional and provincial level) of the dataset. The temporal dimension can be explored thanks to a time slider that also shows the trend of the total number of tweets and the percentage of HS. In figure 5 there's an example of how the choropleth map and the Dorling map appear on June 29, 2019 when the migrant and NGO themes, in a single day, become viral in the public debate[10]. In Figure 4 we see another example: the volume of tweets about the Roma topic in the days from 3 to 5 June 2019 has increased considerably due to some clashes in the outskirts of Rome[11].

[10] http://www.ansa.it/sito/notizie/politica/2019/06/28/sea-watch-indagata-la-capitana.-nuovo-affondo-di-salvini-contro-lolanda-comportamento-disgustoso_991189d6-7818-48d9-b4d8-a2a7d10d31bc.html

[11] http://www.ansa.it/sito/notizie/cronaca/2019/04/04/simone-il-quindicenne-di-torre-maura-contro-casapound-state-a-fa-leva-sulla-rabbia-della-gente.-plauso-raggivideo_7a4bc495-bb4d-4c21-a1f7-2ecbc8422ea5.html

The liquid gauge allows the user to quickly detect the tweet volume increase, from $1,619$ to $14,778$, and the increase in HS rates, from 13% to 23%.

4.2 Words of Hate

Figure 2 shows another visualization: a bar chart containing the 25 words more frequently occurring in the tweets collected in the selected time period. For each word, the user can also see the average percentage of HS in tweets containing that word. As before, the example in figure 2 refers to June 29, 2019. By clicking on a word, the user can visualize additional information about it, such as the exact number of occurrences in the tweets or its co-occurrence network (figure 3).

5 Countering online hate speech in High Schools

The interactive hate maps and the 'Words of Hate' visualization settings described here have been also used within educational paths developed for citizenship and mostly targeting high school students. Such paths were focused on the dismantling of negative stereotypes against immigrants, Roma, and religious minorities, and on the creation of positive narratives to actively counteract hatred online. Since today, a team of twenty educators carried out 90 laboratories in seven different Italian regions (Piedmont, Tuscany, Liguria, Emilia

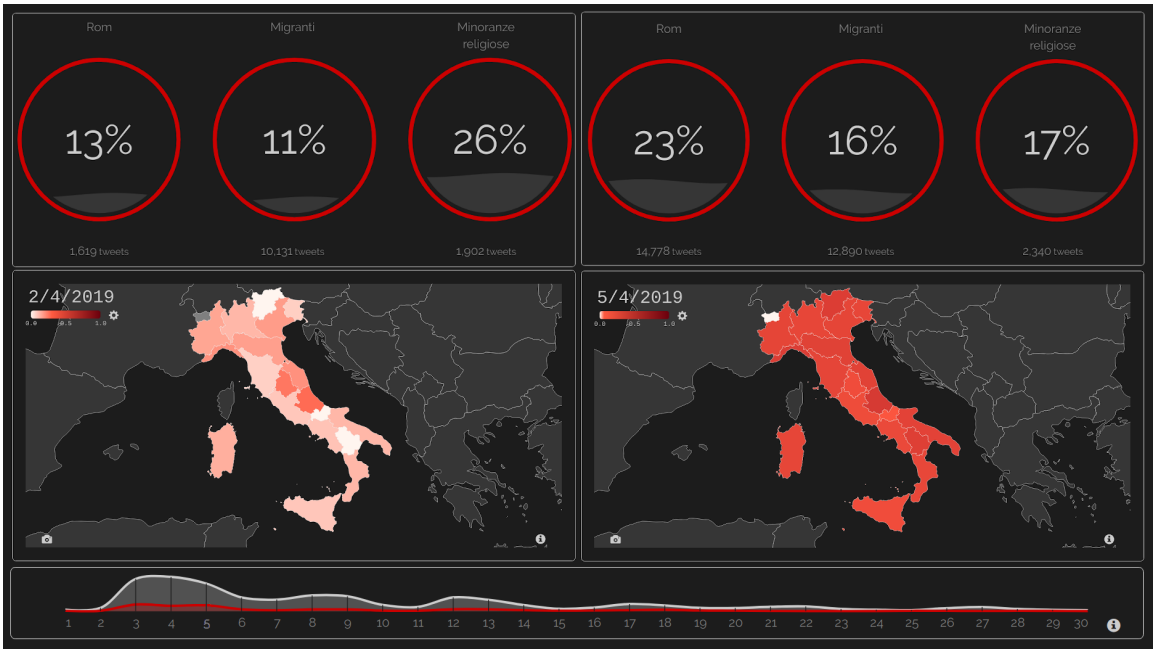

Figure 4: Choropleth map and liquid fill gauge.

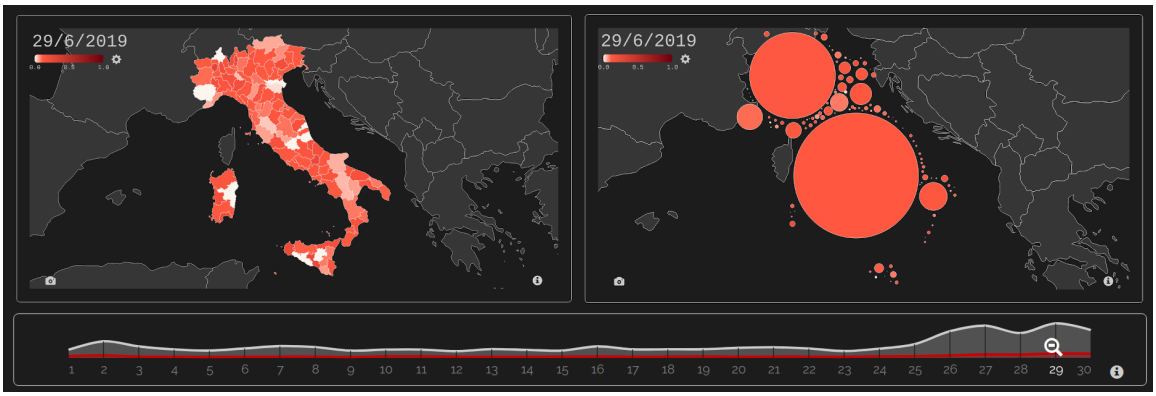

Figure 5: Choropleth map and Dorling map.

Romagna, Lazio, Friuli-Venezia Giulia, and Sardinia). At the end of the project 150 classes will be reached, and the resulting positive narratives will be published on the project website.

6 Conclusion and Future Work

In this paper we described an online platform for monitoring HS against immigrants in Italy at different levels of granularity, which uses Twitter as data source and combines HS detection and advanced visualization techniques in order to provide users with an interactive interface for the exploration of the resulted data. Another important research outcome of the project is HATECHECKER, a tool that automatically detects *hater users* in online social networks, which will be accessible from the platform soon. Given a target user, the workflow that is going to be implemented in our system uses sentiment analysis techniques to identify hate speech posted by the user, and exploits a lexicon-based approach to assign to the person one or more labels that describe the nature of the hate speech she posted (e.g., racism, homophobia, sexism, etc.). A map of Italian projects and associations that spread a culture of tolerance is also under development, to allow 'Contro l'Odio' users to get a better understanding of the HS phenomenon and of the active forces fighting it on the Italian territory.

Acknowledgments

The work of all the authors was partially funded by Italian Ministry of Labor (*Contro l'odio: tecnologie informatiche, percorsi for-*

mativi e storytelling partecipativo per combattere l'intolleranza, avviso n.1/2017 per il finanziamento di iniziative e progetti di rilevanza nazionale ai sensi dell'art. 72 del decreto legislativo 3 luglio 2017, n. 117 - anno 2017).

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63. Association of Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *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.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263, pages 1–9. CEUR.

Tullio De Mauro. 2016. Le parole per ferire. *Internazionale*. 27 settembre 2016.

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)*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *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.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. In *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*, page 214–228. CEUR-WS.org.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2019. Mapping twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, 0(0):1–11.

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the 3rd Workshop on Abusive Language Online, co-located with ACL 2019*. Association of Computational Linguistics.

Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. 2016. Modeling community behavior through semantic analysis of social data: The italian hate map experience. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 307–308. ACM.

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), Rome, Italy, December 11-13, 2017.*, volume 2006 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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)*. CEUR.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics.