Recognizing Hate with NLP: The Teaching Experience of the #DeactivHate Lab in Italian High Schools

Simona Frenda[1,2], **Alessandra Teresa Cignarella**[1,2], **Marco Antonio Stranisci**[1], **Mirko Lai**[1], **Cristina Bosco**[1] **and Viviana Patti**[1]

1. Università degli Studi di Torino, Italy
2. Universitat Politècnica de València, Spain

{simona.frenda|alessandrateresa.cignarella|marcoantonio.stranisci|mirko.lai}@unito.it
{cristina.bosco|viviana.patti}@unito.it

Abstract

The possibility of raising awareness about misbehaviour online, such as hate speech, especially in young generations could help society to reduce their impact, and thus, their consequences. The Computer Science Department of the University of Turin has designed various technologies that support educational projects and activities in this perspective. We implemented an annotation platform for Italian tweets employed in a laboratory called #DeactivHate, specifically designed for secondary school students. The laboratory aims at countering hateful phenomena online and making students aware of technologies that they use on a daily basis. We describe our teaching experience in high schools and the usefulness of the technologies and activities tested.

1 Introduction

Recently, the presence of digital technologies in our lives has grown enormously, with a strong impact on our daily lives. Digital spaces and social media have become a privileged channel for communication, information and socialization, frequented by millions of people at the same time. Along with the new relational opportunities and access to knowledge, even misbehaviour have acquired new visibility and virality, such as hate speech. In spite of a causal link between hate speech and crime is hard to demonstrate, the risk of offences and effects on psychological and physical well-being of the victims are clear in psychological and social studies (Nadal et al., 2014; Fulper et al., 2014). The extreme consequence of these effects might be suicide, especially considering the adolescents, as suggested by recent studies investigating the link between cyberbullying and suicidal behaviors of U.S. youth (Nikolaou, 2017). To prevent such scenarios, few awareness-raising projects in schools are activated by NGOs in Italy, such as Amnesty International[1] or Cifa ONLUS[2].

The *Commissione Orientamento e Informatica nelle scuole*[3] supports a manifold of activities with the main goal of creating a link between schools and academia, also in the context of the national project *Piano Lauree Scientifiche* (PLS). The members of the CCC (Content-Centered Computing) group of the Computer Science Department of the University of Turin, active in the investigation of hate speech online[4], have led and participated in several hate-speech-related projects, including "Contro l'odio"[5] (Capozzi et al., 2020) a joint work with non-profit entities and University of Bari that aims at monitoring hate speech against minorities in Italy. Within the current experience, we created a data annotation platform specifically dedicated to support educational activities and aimed at reflecting on the importance of a conscientious communication. In this perspective, the idea of #DeactivHate takes hold. This laboratory, addressed at students of secondary schools, is articulated in three main modules with the purpose of:

1) raising awareness about this social problem, encouraging the reflection on microaggressions, hate speech, stereotypes, prejudices;

2) stimulating the so-called *computational thinking* and the study of linguistic elements that are exploited by users to offend or to ex-

[1] http://di.unito.it/silencehateitaly.
[2] http://di.unito.it/iorispetto.
[3] http://di.unito.it/orientamentoscuole.
[4] http://hatespeech.di.unito.it/.
[5] https://controlodio.it/.

press hate against a victim online (hashtags, emoticons, or figures of speech);

3) introducing high schoolers to how tools based on NLP (Natural Language Processing) work to incentivize a more conscious use of technology.

To reach these purposes, We designed a series of educational activities that include: analysis of the online problem by means of an investigation on own social networks personal profiles; linguistic analysis of the hateful messages during the annotation of tweets on the "Contro l'odio" annotation platform; manual identification of hate speech in Italian texts playing the role of 'being an automatic classifier'; translation of this task in a real automatic task, coding two types of classifiers in Python. These activities, delivered online due to the pandemic restrictions, have been distributed in 5 meetings (lasting 2 hours each) for each class, between April and June 2021, for a total of 10 hours per class.

2 Related Work

A popular workshop series on the topic of "Teaching NLP" has been recently held on its fifth edition at NAACL-HLT 2021 (Jurgens et al., 2021), where the participants discussed and shared experiences on a variety of important issues such as: teaching guidelines, teaching strategies, adapting to different student audiences, resources for assignments, and course or program design. The main lesson learned has been that of highlighting the importance of creating materials describing NLP, not only for learners at a university/college level, but also for those learners who are younger and have diverse educational backgrounds. In this regard, a great inspiration for starting to work with schools in Italy derives from the experience of Sprugnoli et al. (2018), where the authors – although with different goal in mind than ours – started a project involving NLP and pupils from Italian schools, aged 12-13. That experience was chiefly dedicated to the study of cyberbullying among pre-teens and the creation of a corpus of WhatsApp threads in the context of the CybeRbullying EffEcts Prevention activities (CREEP) project. Our idea of starting a project that could bring NLP to high schoolers and that, at the same time, could introduce the themes of hate speech, microaggressions, and discrimination by eliciting personal experiences and students' opinions, is somehow in continuity with that experience.

A second work of great relevance for the creation of our experience, has been the reading of Pannitto et al. (2021), in which the authors point out, for the first time, the fact that no high school curricula in Italy includes any *(computational) linguistics* education and that the lack of this kind of exposure makes choosing computational linguistics as a university degree unlikely. Furthermore, the authors highlight that NLP is, indeed, at the core of many tools young people use in their everyday life, and having almost zero knowledge of this field makes the use of such tools less responsible than it could be. The authors have been the first to create a dedicated workshop for Italian, aimed at raising awareness of Italian students aged between 13 and 18 years regarding the subject of NLP (Messina et al., 2021).

Additionally, the idea of creating some playful and meaningful activities regarding NLP and the themes of hate speech for high schoolers, are in line with the concept of '*gamification*', which lately has been applied to many linguistic annotation tasks, as an alternative to crowdsourcing platforms to collect annotated data in an inexpensive way (Bonetti and Tonelli, 2020), such as our "Contro l'odio" annotation platform.

3 #DeactivHate

The goals of #DeactivHate are: 1) raising awareness about misbehaviour online, such as hate speech, eliciting also personal experiences, 2) stimulating computational thinking and linguistic observation of hateful messages, and 3) encouraging a conscious use of technologies discovering how they work. To reach these objectives we articulated three modules as described below.

3.1 Hate Speech: Introduction

The first module aims at introducing a definition of hate speech to students. Hate speech is often mistaken for a generic insult rather than a specific phenomenon "connected with hatred of members of groups or classes of persons identified by certain ascriptive characteristics (e.g., race, ethnicity, nationality)" (Brown, 2015).

The session started with an ice-breaking activity in which students presented themselves through an image found online, depicting an aspect of their identity (see Figure 1). We then asked them to tell whether they were ever attacked or stigmatized for this characteristic.

Figure 1: Example of Jamboard of Google

In this way, we guided the class in drawing a distinction between **non-ascriptive** identity traits (e.g., political belief, style of dressing) and **ascriptive**[6] ones (e.g., ethnicity, sexual orientation, skin colour) (Reskin, 2005). The idea behind this activity is twofold: i) it links issues such as hate speech and racial microaggression (Sue, 2010) to students' lives; ii) it helps distinguishing the spreading of discriminatory contents[7] from generic insults. The module ended with an assignment: students had to find at least one public figure who had been a victim of online discrimination, providing one or more hateful messages as an example, and a counter-narrative response.

3.2 "If I Were a Classifier..."

The second module is organized in two meetings and focuses on the importance of manually annotated corpora for online hate speech detection and what are the peculiarities of hateful messages.

Within the first meeting, each student presented the found messages and try to define the type of attack and the linguistic characteristics of the text that make it hateful or a counter-narrative. The variety of examples led to the introduction of a deeper taxonomy of discrimination (e.g., misogyny, homophobia, sexism, etc...). As expected, the following group discussion brought out a considerable subjectivity in perceiving these phenomena, thus highlighting the need of adopting a shared annotation schema to identify hate speech in messages.

After a brief introduction on what corpora are and how they are used in new technologies, students have been involved in an annotation task of hate speech, asking them to evaluate at least 30 tweets.

For this purpose, we created the data annotation platform[8] within the "Contro l'odio" project. This web application, built using PHP, MySQL, and JavaScript, [9], preserves the student's annotation history by using a passwordless authentication link sent to the email chosen during the login. This method has the twofold advantage of not requiring the student to register to the platform and of preventing ourselves to save the student's email or other personal data. It then ensures the annotation anonymity and satisfies the requirements of General Data Protection Regulation (GDPR), as a desired consequence.

The home page of the web application consists of a dashboard that provides the annotation guideline and shows basic information about the student's activity. Indeed, the student could know the number of sessions they completed (each session consists of annotating 15 tweets) and the level of agreement (expressed in percentage) between their annotation and the annotation performed by the automatic model realized in the "Contro l'odio" project. Gamifying the task through this comparison, we provide the basis for a discussion about the fallibility of automatic systems. Furthermore, we also allow the student to compare their annotation with the annotation of their classmates in order to introduce the measures of annotator agreement. When a session starts, the student could annotate the level of hatefulness of a tweet through a 7 square scale filled with a color scale from *Watusi* to *Sangria* as shown in Figure 2. Two additional squares, respectively filled with *White* and *Mid-Gray*, allow stating the absence of hate or to consider off-topic the content of the tweet. Finally, three toggle switches (on/off button) were added to check the presence of 'irony/sarcasm/humor', 'offensiveness', and 'stereotype', giving them the possibility to reflect about the ways in which users spread hate online.

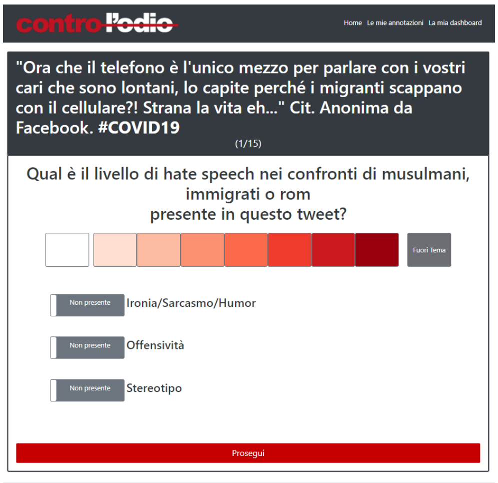
During the annotation task, students were asked to fill a shared spreadsheet with the tweets that impressed them the most for its offensiveness, for its humorous intention, or the most difficult to anno-

[6] Qualities beyond the control of an individual.

[7] The definition of hate speech we referred to is the one codified by The Council of Europe: "the term 'hate speech' shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance" (Recommendation No. R (97) 20).

[8] https://didattica.controlodio.it/.

[9] https://github.com/mirkolai/DEACTIVHATELab.



"Ora che il telefono è l'unico mezzo per parlare con i vostri cari che sono lontani, lo capite perché i migranti scappano con il cellulare?! Strana la vita eh..." Cit. Anonima da Facebook. #COVID19

(1/15)

Qual è il livello di hate speech nei confronti di musulmani, immigrati o rom presente in questo tweet?

Ironia/Sarcasmo/Humor

Offensività

Stereotipo

Figure 2: Data Annotation Platform

tate. By discussing with them annotation results, we introduced the latest core concept of the module: the **agreement**. We presented some metrics that are typically adopted to calculate it among annotators and outlined some good practice recently emerged in Corpus Linguistics, such as ensuring the involvement of minorities in corpora development in order to avoid biases (Basile, 2020).

3.3 My First Classifier

In this module the main idea is to stimulate computational thinking by translating linguistic observations coming from the annotation procedure in a proper computational task. The activity of annotation has, indeed, given the opportunity of reflecting on how users tend to verbally express hate online, and on how minorities are represented through stereotypes. To incentivize this transition, we proposed two activities:

A. to mark in each tweet the textual span that could make a classifier aware of the presence of hate speech creating a list of word n-grams;

B. to develop two automatic classifiers (supervised and unsupervised) exploiting the list of word n-grams.

Before starting with the first activity, we asked students to motivate their choice of the tweets selected during the previous exercise. Some tweets triggered a discussions on what should be considered hate speech or not, and the doubts were later solved by looking at the provided definitions of hate speech and at the annotation guidelines. The most controversial tweets report aggressive events or racial propositions; and, for this reason, they were perceived as hurtful by the majority of the students:

(i) *Autobus per i bianchi e altri per i migranti. Non si parla dell'apertheid del Sudafrica né del periodo di segregazione negli Stati Uniti, ma di una proposta della Lega per la provincia di Bergamo. L'Italia non è un paese razzista ma nel 2020 questo è ciò di cui si discute. URL*[10]

Others triggered interesting linguistic reflections, such as:

(ii) *Peccato che non sbarcano povere famiglie africane, ma solo mafia nigeriana, ex galeotti tunisini, stupratori senegalesi, terroristi dell'Isis dalla libia, tutti criminali robusti 1.80 di altezza, pronti a spacciare droga, violentare le nostre donne, cannibali e assassini.*[11]

In these, the students retrieved specific figures of speech such as sarcasm, rhetorical questions and analogies, and also strong words that reflect the social biases towards the minorities. In activity A, all the words and expressions that could make the message hurtful have been collected in a list of n-grams of words called `our_lexicon` (Table 1). Following, the items of such list have been exploited by the classifiers to predict if a tweet contains hate speech or not.

unigrams	risorse, sporchi, pacchia, schifo, invasione, spacciare
n-grams	porti chiusi, cacciarli via, difesa della patria[12]

Table 1: Examples from `our_lexicon`

For activity B, we created an interactive Python notebook using the *Colaboratory* platform provided by Google, as a similar initiative had successfully been carried out by Hiippala (2021) with a similar educational tool. To allow the students to use the notebook in spite of their computer skills, we elaborated some guidelines explaining even how to create a folder in Google Drive and

[10] Translation: *Buses for whites and others for migrants. There is no mention of South Africa's apartheid or the period of segregation in the United States, but of proposal by Lega for the province of Bergamo. Italy is not a racist country but in 2020 this is what we are discussing. URL.*

[11] Translation: *Too bad that poor African families do not land, but only the Nigerian mafia, former Tunisian convicts, Senegalese rapists, ISIS terrorists from Libya, all heavy-weight criminals 1.80 tall, ready to sell drugs, rape our women, cannibals and murderers.*

[12] Translation: Unigrams: resources, dirty, *godsend*, disgust, invasion, peddle. N-grams: closed harbours, send [them] away, defence of the fatherland.

how to import all the necessary materials inside of it. Among the required materials, we prepared the dataset using the tweets previously annotated by the students.

We proposed two types of classifiers:

1) unsupervised classifier based on the list `our_lexicon` for which if one of the selected grams are inside the text, the text is predicted as hateful;

2) supervised classifier based on Support Vector Machine algorithm using the list `our_lexicon` as main feature of the classification task.

The coding of the first classifier allowed students to gain confidence with some basics of Python; whereas the second one introduced them to core of new technologies based on machine learning (see Figure 3). At the end of the activity, we observed together the performances of automatic systems and analyzed some of the tweets that were wrongly classified. This final step helped students to reflect on the limitations of machines and the important role of the linguistics in language-related technologies.

4 What We Learnt

Due to pandemic restrictions, we taught the entire laboratory through remote modality (DAD)[13] between April and June 2021 to 2 classes of one secondary school of Turin, with students aged 16-20. As described above, various resources and tools have been used (and created *ex novo*) to bring forward the educational activities in distant teaching mode. However, we plan to propose the same activities/materials even for lessons *in presentia* exploiting the computer rooms of the schools.

For each class, we organized the activities of the three modules in 5 meetings of about 2 hours. Despite the shortness of the laboratory, we found that realizing specific activities for each session helped us manage efficiently the available time. We resorted to web applications to make up for the different devices and operating systems used by the students at their homes. And, in particular, we used Google Meet, as it offers interactive tools such as virtual blackboard, and Moodle, a learning platform provided by the University of Turin that gave us the possibility to organize our activities making available the necessary materials to students. Moreover, each meeting was supported by the use of slides for having visual and descriptive support. The classes assisted in this short period were composed of a total of 35 adolescents, coming from different countries. From the first meeting they showed a general interest in the treated subject, and we were surprised especially by the profoundness of some observations raised during the discussions. The students, indeed, were encouraged to share their opinions, doubts, and perspectives. These discussions made clear that the students face these problems related to technology and communication every day, sometimes suffering even the consequences. Hate speech is, indeed, a very sensitive issue and the perception of what is abusive or not, depends on the cultural background of each student. This fact, on the one side stimulated the debates, however, on the other side, it made it difficult for us to find the *ideal* way to share complex concepts and manage specific situations.

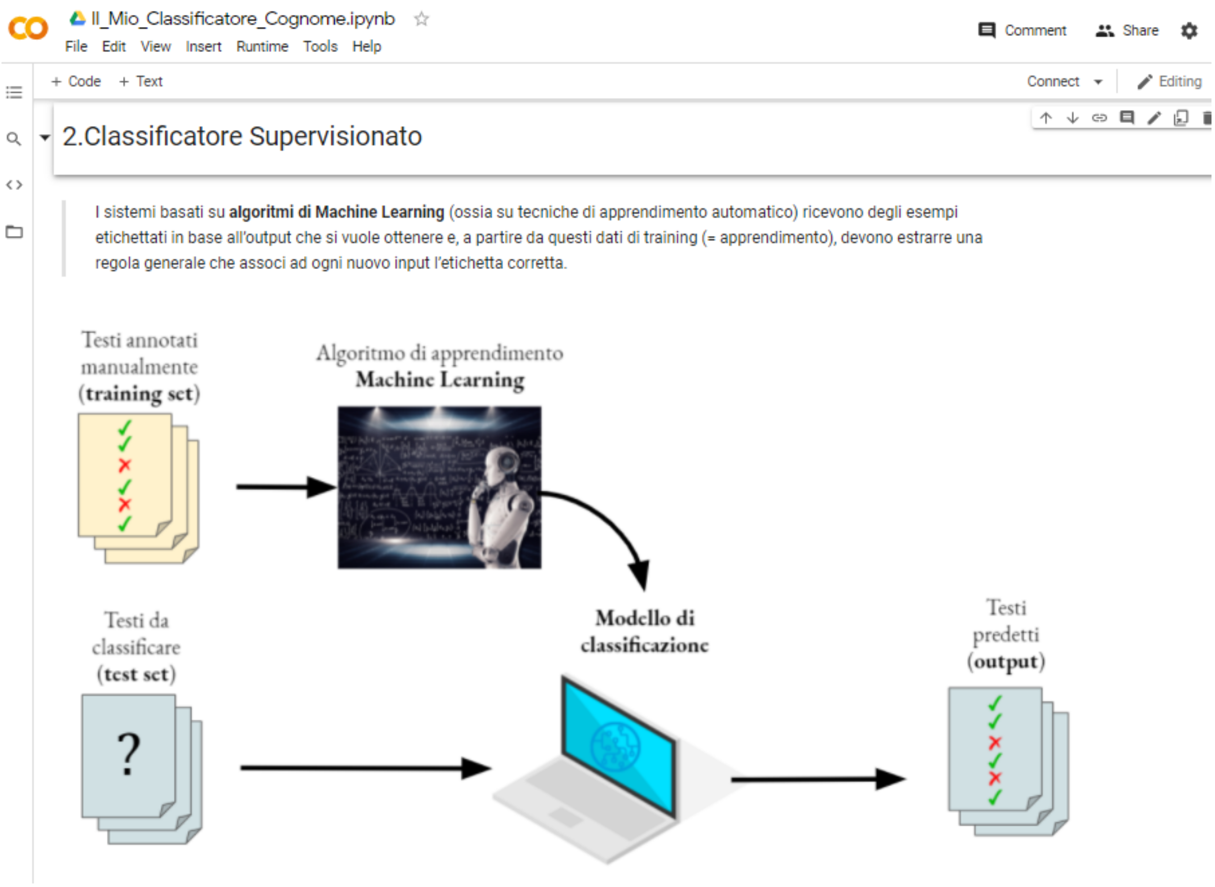
At the end of the laboratory, we provided a survey in order to collect the impressions and the opinions of students. Analyzing these surveys, we noticed that the majority of students considered interesting the content of #DeactivHate, but it appears clear that the format online of the laboratory was perceived from students less interactive and fluent, due especially to technical problems when a part of students were in class and other part at home[14]. From our perspective, we noticed an interesting difference between younger and older students. The older were more active during the activities and discussions than the younger. Moreover, we thought that the number of students affected the flow of the debates, especially in the DAD context. We expect that *in presentia* the proposed activities could have a better impact facilitating the interaction.

5 Conclusion

#DeactivHate represents for Italian high schoolers a first step towards the introduction to subjects such as Linguistics and NLP, that are, for the most part, unknown in Italian high schools, in spite of their relevance in everyday technology. Indeed, this kind of laboratory reveals what are the possible hybrid and multidisciplinary applications

[13] Didattica A Distanza.

[14] For the most part of the school year 2020-2021, Italian schools allowed a capacity of 50% inside classrooms.


Il_Mio_Classificatore_Cognome.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
Connect
Editing
Testi annotati manualmente (training set)
Algoritmo di apprendimento Machine Learning
Modello di classificazione
Testi da classificare (test set)
?
Testi predetti (output)

2.Classificatore Supervisionato

I sistemi basati su **algoritmi di Machine Learning** (ossia su tecniche di apprendimento automatico) ricevono degli esempi etichettati in base all'output che si vuole ottenere e, a partire da questi dati di training (= apprendimento), devono estrarre una regola generale che associ ad ogni nuovo input l'etichetta corretta.

Figure 3: Supervised Classifier Section on Python Notebook

of Computer Science and Linguistics related degrees, far from the *conventional* employment opportunities. Looking at the future, we would like to enhance the proposed activities in order to make them more interactive even in an online context (such as the DAD) following the example of Hiippala (2021).

A final remark needs to be made regarding the lack of evaluative strategies that could allow us to understand the impact of #DEACTIVHATE in students' online behaviors or their knowledge of technologies. Therefore, following the example of Bioglio et al. (2018) and Athanasiades et al. (2015), in the next editions we have planned to employ: surveys before and after the intervention to evaluate the online activity of the students and their experiences about misbehavior (caused or suffered); and interviews to teachers after the conclusion of the laboratory to understand if some changes were perceived with respect to the class group. Future activities will integrate also basic evaluations to assess the degree of learning with respect to the contents of the course, such as computational thinking, annotation methodologies, automatic text processing, as well as a final evaluation of the proposed teaching activities collecting the personal impressions of the students.

In addition, to validate also the impact of #DEACTIVHATE in the society and, in particular, in the city context we think to measure the detection of the amount of hateful message online by means of monitoring platforms, such as the "Contro l'odio" map.[15]

Acknowledgements

The work of S. Frenda, A. T. Cignarella and M. Lai has been funded under the national project *Piano Lauree Scientifiche (PLS) 2019/20* as part of the activities of *Computer Science Department, School of Science of Nature, University of Turin*. The authors would like to extend a special thanks to the school 'Convitto Nazionale Umberto I', and in particular, to Professor Simona Ventura for her availability and her collaboration in this adventure with #DEACTIVHATE.

[15] `https://mappa.controlodio.it/`.

References

Christina Athanasiades, Harris Kamariotis, Anastasia Psalti, Anna C Baldry, and Anna Sorrentino. 2015. Internet use and cyberbullying among adolescent students in Greece: the “Tabby” project. *Hellenic Journal of Psychology*, 12(1):14–39.

Valerio Basile. 2020. It's the End of the Gold Standard as We Know It. In *International Conference of the Italian Association for Artificial Intelligence*, pages 441–453. Springer.

Livio Bioglio, Sara Capecchi, Federico Peiretti, Dennis Sayed, Antonella Torasso, and Ruggero G Pensa. 2018. A social network simulation game to raise awareness of privacy among school children. *IEEE Transactions on Learning Technologies*, 12(4):456–469.

Federico Bonetti and Sara Tonelli. 2020. A 3D Role-Playing Game for Abusive Language Annotation. In *Workshop on Games and Natural Language Processing*, pages 39–43. European Language Resources Association.

Alexander Brown. 2015. *Hate speech law: a philosophical examination*. Routledge.

Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, et al. 2020. “Contro L'Odio”: A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media. *IJCoL. Italian Journal of Computational Linguistics*, 6(6-1):77–97.

Rachael Fulper, Giovanni Luca Ciampaglia, Emilio Ferrara, Y Ahn, Alessandro Flammini, Filippo Menczer, Bryce Lewis, and Kehontas Rowe. 2014. Misogynistic language on Twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on ChASM*.

Tuomo Hiippala. 2021. Applied Language Technology: NLP for the Humanities. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 46–48, Online, June. Association for Computational Linguistics.

David Jurgens, Varada Kolhatkar, Lucy Li, Margot Mieskes, and Ted Pedersen, editors. 2021. *Proceedings of the Fifth Workshop on Teaching NLP*. Association for Computational Linguistics.

Lucio Messina, Lucia Busso, Claudia Roberta Combei, Alessio Miaschi, Ludovica Pannitto, Gabriele Sarti, and Malvina Nissim. 2021. A Dissemination Workshop for Introducing Young Italian Students to NLP. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 52–54, Online, June. Association for Computational Linguistics.

Kevin L Nadal, Katie E Griffin, Yinglee Wong, Sahran Hamit, and Morgan Rasmus. 2014. The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling & Development*, 92(1):57–66.

Dimitrios Nikolaou. 2017. Does cyberbullying impact youth suicidal behaviors? *Journal of health economics*, 56:30–46.

Ludovica Pannitto, Lucia Busso, Claudia Roberta Combei, Lucio Messina, Alessio Miaschi, Gabriele Sarti, and Malvina Nissim. 2021. Teaching NLP with Bracelets and Restaurant Menus: An Interactive Workshop for Italian Students. In *Proceedings of the Fifth Workshop on Teaching NLP*, Online. Association for Computational Linguistics.

Barbara F. Reskin. 2005. Including mechanisms in our models of ascriptive inequality. *Handbook of employment discrimination research*, pages 75–97.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59. Association for Computational Linguistics.

Derald Wing Sue. 2010. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons.