# From Hate Speech to Societal Empowerment: A Pedagogical Journey Through Computational Thinking and NLP for High School Students

**Alessandra Teresa Cignarella**[△,♡], **Elisa Chierchiello**[★], **Chiara Ferrando**[★], **Simona Frenda**[★,♡], **Soda Marem Lo**[★] and **Andrea Marra**[★]
★ Computer Science Department, University of Turin, Italy
△ LT3, Language and Translation Technology Team, Ghent University, Belgium
♡ aequa-tech, Turin, Italy

## Abstract

The teaching laboratory we have created integrates methodologies to address the topic of hate speech on social media among students while fostering computational thinking and AI education for societal impact. We provide a foundational understanding of hate speech and introduce computational concepts using matrices, bag of words, and practical exercises in platforms like Colaboratory. Additionally, we emphasize the application of AI, particularly in NLP, to address real-world challenges. Through retrospective evaluation, we assess the efficacy of our approach, aiming to empower students as proactive contributors to societal betterment. With this paper we present an overview of the laboratory's structure, the primary materials used, and insights gleaned from six editions conducted to the present date.

**Our positionality:** This paper is situated in (Northern) Italy in 2024 and is authored by researchers specializing in Natural Language Processing. Beyond our academic work, we are actively involved in *feminist*, *LGBTQIA+ advocacy*, and *anti-hate speech* activism. Collectively, our backgrounds span theoretical linguistics, computer science, natural language processing, digital humanities, high school teaching, and non-formal education methodos.

## 1 Introduction

The pervasive use of technologies based on AI models, makes it imperative for academic institutions to organize teaching laboratories for primary and secondary schools with the aim of increasing awareness for the techniques behind these technologies, consequently knowing when to trust AI and when to distrust it, and revealing the "behind the scenes" of their unconscious use. Some programs are promoted also by government institutions with the aim of bringing students closer to computer science and also reducing the gender stereotypes that characterize this field of study. Among them, are worth mentioning: Women Who Code, supported by the EU, and Coding Girls in Italy.

In this context of activities for public engagement, our laboratory **#DEACTIVHATE** takes shape. Its main goals are: introducing secondary school students to Natural Language Processing techniques and their applications; raising awareness about the ethical issues of digital world; empower them to positively contribute to the digital community, and increase responsibility in the use of present-day technologies.

To achieve these goals, we designed a series of educational activities starting from the analysis of online hate speech. Abusive and online harmful content are issues that adolescents face in their everyday life, but also one of the social issues that they can help alleviate. In spite of a causal link between hate speech and crime is difficult to prove, the risk of offenses and effects on victim's psychological and physical well-being have been proved in psychological and social studies (Nadal et al., 2014; Fulper et al., 2014). Especially among adolescents, the extreme consequences of these attacks tend to be the suicide, as suggested by (Nikolaou, 2017) in their analysis of the connection between cyberbullying and suicidal behavior in the US. To prevent such scenarios, some awareness-raising projects in schools are being carried out by NGOs in Italy, such as Amnesty International[1] or Cifa ONLUS[2]. #DEACTIVHATE fits in this context, merging the educational experience of development and use of AI-based tools and the stimuli to be responsible developers and users.

Our experience of teaching this laboratory concerned students of different ages and coming from different backgrounds: humanistic, classical, technical and scientific studies. Therefore, the methodologies of teaching used in this context, and the

[1] https://www.silencehate.it/.
[2] https://www.cifaong.it.



*Proceedings of the Sixth Workshop on Teaching NLP*, pages 54–65

materials and activities employed during the laboratory, are adaptable to different situations.

The impact of the laboratory has been evaluated by administering tests in two phases: one at the beginning and one at the end, containing an (almost) identical set of questions. By means of these pre- and post-test we could assess the teaching methodologies and materials and to measure the awareness of students towards: firstly, the functionality of AI-based technologies, secondly, the importance of creating responsible and ethical NLP for community benefits, and thirdly, the consequences of pervasive online hate speech.

In the next sections, we describe: related work on teaching NLP that report experiences with young participants (Section 2); the methodologies and teaching activities and materials employed in our laboratory (Sections 3 and 4); our experience with different Italian secondary school students (Section 5). Finally, we write about some of the challenges we faced, and we delineate some conclusions (Sections 6 and 7).

## 2 Related Work

The escalation of hate speech on social media platforms and its negative societal impact have ignited significant academic interest in developing methods for its automatic detection and mitigation. This surge in research is underscored by the proliferation of methodologies leveraging Natural Language Processing and Machine Learning (ML) techniques. A comprehensive survey (Jahan and Oussalah, 2023) delineates the evolution of automatic hate speech detection, emphasizing the integral role of NLP and Deep Learning (DL) technologies in this realm. Their systematic review delineates the progression from traditional ML techniques to advanced DL architectures, highlighting a shift towards models like BERT, which have revolutionized hate speech detection with their context-aware processing.

In parallel, educational initiatives have emerged as critical for cultivating a responsible digital citizenry, particularly among the younger population. This educational aspect aligns with our project's dual focus: addressing hate speech through technological solutions, while promoting computational thinking and AI literacy among students. Workshops like the one discussed at NAACL-HLT (Jurgens et al., 2021) emphasize the importance of developing NLP resources for diverse educational contexts, reflecting the necessity of embedding these technological competencies at an early age. Furthermore, other initiatives (Sprugnoli et al., 2018; Pannitto et al., 2021) illustrate the emerging trend of integrating computational linguistics into the high school curriculum, thereby aligning with our laboratory's educational objectives.

Our approach to combating hate speech incorporates practical exercises and the utilization of platforms such as Colaboratory, fostering an environment where students not only learn to identify and counteract hate speech but also gain hands-on experience with NLP tools. This pedagogical strategy mirrors the "gamification" techniques highlighted by Bonetti and Tonelli (2020), which have been effectively applied in linguistic annotation tasks, enhancing engagement and educational outcomes.

Reflecting on the systematic review and related educational efforts, our project's methodology synthesizes these insights, employing state-of-the-art NLP techniques for real-world applications while fostering an educational paradigm that prepares students to navigate and contribute positively to the digital world.

### 2.1 A bit of History

The #DEACTIVHATE project was conceived by a group of young researchers within the initiatives for the orientation of high school students and in particular for the promotion of STEM subjects among the young female population. It is supported by *Commissione Orientamento e Informatica nelle Scuole* and funded by the project "Piano Lauree Scientifiche" of the Computer Science Department in the University of Turin.

Through the six editions of the lab, 14 different classes were reached, for a total number of 233 students, aged from 15 to 18 years old (see Table 1 in Appendix A). The first two editions involved students with a humanistic background, while in the following ones students from technical or scientific high schools – thus with a stronger background in computer science – were reached. The results of the first three editions of the lab were discussed in (Frenda et al., 2021; Cignarella et al., 2023).

With the present publication, we aim at describing the hands-on experience of the three new (post-COVID) editions. In particular, here we tackle most of the issues raised in the "Future Work" sections of previous publications. Some have been resolved or confirmed, while others remained open and are due to further discussion with the teaching

community. For example, there was a request to make the lab more interactive in its online setting, or to expand the lab beyond the context of Turin, which happened with the fifth edition (even if only online). Furthermore, we found it crucial to present #DEACTIVHATE in a new, more comprehensive publication. After six editions, the laboratory has evolved into a well-refined and effective program.

In addition, we provide an in-depth description of the materials developed for the laboratory, we translated all of them into English, making them accessible to a wider and international audience (see Appendix B). Finally, acknowledging the various limitations our laboratory may still have, we expect to receive feedback from the teaching community and that #DEACTIVHATE will be adopted in new schools and different contexts.

## 3 Teaching Goals and Methodologies

The laboratory's name, #DEACTIVHATE, combines the concept of deactivation with the phenomenon of hate, and the new term is preceded by the pound sign '#', reminiscent of social media hashtags. This choice wants to establish a clear connection to the social media realm. The activities, designed for secondary school students, consist of three main modules aimed at:

1. Raising awareness about the pervasive issue of hate speech, prompting reflection on microaggressions, stereotypes, and prejudices.
2. Engaging students in computational thinking and exploring linguistic tools used by social media users to convey hate or offend others online, such as hashtags, emoticons, and rhetorical devices.
3. Introducing high school students to Natural Language Processing (NLP) tools and demonstrating their potential for promoting more responsible and conscious technology usage.

By combining educational content with hands-on exploration and critical thinking exercises, #DEACTIVHATE strives to empower students to become discerning and empathetic digital citizens.

In order to achieve these purposes, we relied on the following methodologies:

• **Collaborative reading sessions:** Students engage in reading formal definitions and in exploring the basics of hate speech, including vocabulary and definitions provided by authoritative sources such as the Council of Europe.

• **Matrix design and analysis:** Utilizing Google Spreadsheets, students design matrices incorporating binary (0s and 1s) representations of keywords and concepts, employing techniques such as bag of words to analyze text data.

• **Practical coding exercises:** Students work on exercises using Google Colaboratory, with some exercises pre-compiled and others involving collaborative coding sessions where code is written together to explore concepts related to hate speech detection in NLP.

• **Real-life scenario exploration (Social Media)**: Students engage in browsing social media to gain insight into real-life demonstrations of hateful behaviors and patterns. This activity allows for first-hand exploration of how hatred can manifest on social media platforms and the role NLP plays in identifying, analyzing, and potentially mitigating its effects. By observing and discussing examples from social media, students develop a deeper understanding of the practical implications of NLP in addressing hate speech and promoting responsible online behavior.

## 4 Activities and Materials in Detail

In this section, we describe the teaching activities and the materials employed in #DEACTIVHATE, which are available at the following link: `https://github.com/deactivhate`. The topics of the following 5 lessons cover various disciplines, useful for enhancing knowledge of high schoolers, including: *Sociology/Civics and Hate Speech*, *Computational Linguistics* and *Computer Science/Programming*. For an exhaustive list of the materials, please refer to Appendix B.

### 4.1 Lesson 1: Who are we? Why are we here?

In the first minutes of the first lesson, we administered a pre-test. In order not to "start off with the wrong foot" with the students, we clarified multiple times that the test is designed to assess their pre-existing knowledge on the topics dealt with in the laboratory (and absolutely not for evaluation).

The first lesson sets out to introduce ourselves as university researchers, explain what we do in our research, and set together the overarching goals of the entire laboratory. Students are guided into an introspective and comparative analysis of their own identity. Using Google's Jamboard as a tool, we embark on a journey of self-reflection through an

engaging ice-breaking activity. They are encouraged to present an aspect of their identity using an image, which they upload to a shared Jamboard. Utilizing this tool allows for real-time collaboration and discussion, enriching the learning experience by visually capturing the mosaic of student identities and favoring an environment of empathy and understanding.

Here, students start exploring the multifaceted nature of personal identity, engaging in an introductory dialogue about discrimination and Hate Speech.

*Lesson 1 in brief:* pre-test, icebreaker activity, introductory slides to #DEACTIVHATE.

### 4.2 Lesson 2: How to recognize hate speech?

The second lesson delves into clarifying the concepts introduced in the previous meeting. An initial exploration regarding personal and social identities is proposed, by incorporating the visual tools of the **Pyramid of Hate** and of the **Wheel of Privilege** (see Figure 1) into the slides.

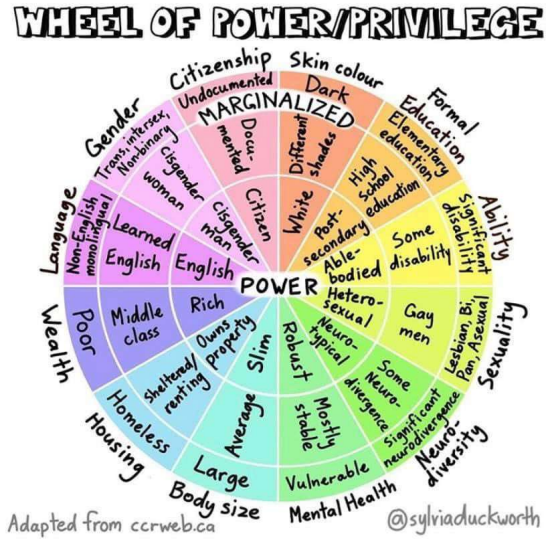


Figure 1: Credits to Sylvia Duckworth.

Thanks to these visualizations, students are encouraged to reflect on their positions within the spectrum of these characteristics, sparking conversation about the relative advantages and disadvantages that accompany different identity markers such as, gender, skin color, body size, wealth etc.[3]

Drawing upon the personal narratives and experiences of the students, the activity culminates in an exploration of the intersectionality of **Hate Speech** with stereotypes, biases and microaggressions, underscoring the nature of such interactions in the fabric of everyday life.

In the second part of this lesson, we introduce the basic terminology to discuss hate speech and related phenomena, relying on the official definition provided by European Commission against Racism and Intolerance (ECRI). This activity sets the stage for a deeper understanding of the phenomenon, emphasizing its targeted nature against identifiable groups based on inherent traits.

> *Hate Speech is to be understood as the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of "race", color, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status [...]*

In the third part of this lesson, students were instructed to open any of the social media accounts they use on a daily basis and try to collect tweets containing hatred messages towards public figures as targets of discrimination.

To organize the analysis and group discussion of their discoveries, they were asked to collect the **textual messages into a Google Spreadsheet**. This method prompted them to identify the keywords in the hateful message, the victim, and categorize the types of discrimination including misogyny, homophobia, sexism, body-shaming, and more, introducing students to a nuanced taxonomy.

As the final activity in this lesson, students are immersed in a hands-on annotation task, where they are asked to analyze and annotate a minimum of 30 tweets. This exercise can be done alone or in pairs or small groups, encouraging discussion, and in our case is facilitated by the tailor-made data annotation platform[4] developed for the project **"Contro l'odio"**. Any other annotation platform can be used.

[3] Both visual materials were developed within the U.S., therefore we adapted them to our needs by, for instance, substituting *English* with *Italian* in the Language section of the Wheel of Privilege. The absence of the topic *Religion* was also noted from the students. Other adjustments might be necessary, depending on the context in which this lab is delivered.

[4] http://annotazione.didattica.controlodio.it/

The lesson is wrapped-up by means of a collective discussion, allowing students to share insights and reflect on the complexities of an annotation task, understanding all the nuances of hate speech and finding an agreement.

*Lesson 2 in brief:* personal and social identity, pyramid of hatred, Hate Speech definition, activity on social media, annotation exercise.

### 4.3 Lesson 3: Machine Learning and matrices

In the third lesson, we introduce students to the fundamentals of machine learning, starting with a broad overview of what it entails and moving into the specifics of **supervised and unsupervised learning**, with a significant focus on the process of text vectorization and the specifics of detecting hate speech through automatic text classification. This lesson is designed to guide students through the entire **machine learning workflow** in the context of NLP. This includes defining a clear task, gathering a suitable dataset, and dividing it into annotated training and test sets.

After the more theoretical aspects, introduced thanks to two sets of slides, the module transitions into a practical activity where students applied their newly acquired knowledge of text vectorization. Each student was tasked with annotating a specific tweet, chosen to reflect the varying types of discriminatory language found online. The activity involved constructing a **bag of words matrix on a Spreadsheet**, where students encoded the presence or absence of certain key terms—terms indicative of the underlying sentiment or hate speech within the tweet. The *one-hot encoding* matrix was used as device to transform the qualitative aspects of language into a quantitative format that machine learning algorithms could process. The same matrix will be created automatically in the coding part of the course (in Lesson 5). By breaking down tweets into this bag-of-words model, students not only practiced the procedure of vectorization but also engaged with the content at a deeper level, considering how individual words contribute to the overall message and tone of the text.

Finally, we used any spare time at the end of this lesson to make sure to install Google Colaboratory and be ready for the next lesson.

*Lesson 3 in brief:* machine learning workflow, supervised/unsupervised learning, training/test set features, vectorization, bag-of-words matrices.

### 4.4 Lesson 4: Python and Colab as IDE

The fourth lesson guides students through the essentials of **Python**[5] programming within the interactive environment of **Google Colaboratory**[6]. The session begins with an overview of Python's basic constructs, using simple print statements to demonstrate output on the screen. This introduction quickly progresses to exercises involving string manipulation, arithmetic operations, and gathering user input—all through the lens of Colab's user-friendly interface.

As the lesson unfolds, students tackle more advanced topics, including string operations and text processing, which are fundamental to NLP tasks. They learn to clean text data, manage strings, and explore the foundational technique of tokenization—turning streams of text into analyzable components. This hands-on experience not only solidifies their Python skills but also prepares them for the subsequent lesson on text classification in NLP.

*Lesson 4 in brief:* Colaboratory, Python, tokenization, lemmatization, word distribution and relevance, n-grams, basic operations with strings.

### 4.5 Lesson 5: Supervised classification

Lesson 5 involves a practical exercise using Colab for detecting hate speech in a given dataset (in our case, sampled from HaSpeeDe2 (Sanguinetti et al., 2018), the benchmark for HS detection in Italian) via a simplified pipeline based on supervised classification.[7] The session begins by defining hate speech in the context of machine learning, utilizing a dataset of tweets categorized by the presence or absence of hate speech (encoded with 0s and 1s).

Students learn to use **pandas**[8] for handling data frames, visualize data, and prepare it for analysis, including balancing the dataset, converting string labels to numerical formats, and splitting data into training and test sets. They also employ text vectorization methods like `CountVectorizer` (bag of words) and `TfidfVectorizer` (words weighted with TF-IDF) from **scikit-learn**[9] to process tweet data for machine learning, as they already have done manually in Lesson 3. During

[5] https://www.python.org/

[6] https://colab.research.google.com/

[7] The dataset used inside this interactive notebook contains Italian texts. Datasets in other languages and on different topics can be found, for instance here: https://live.european-language-grid.eu/

[8] https://pandas.pydata.org/

[9] https://scikit-learn.org/stable/

the lesson, students have the possibility to "play" with the parameters of the `CountVectorizer` and `TfidfVectorizer` methods and select the best textual representation. With the foundation set, they are guided through the construction of a *Support Vector Machine (SVM)* model, applying it to classify tweets and evaluate the model's performance through accuracy metrics. They critically analyze misclassified texts and consider strategies for improving the model, discussing preprocessing functions and the importance of cleaning text data. Based on the students' proficiency with Python and Colab, the class can be guided step-by-step, allowed to work more independently, or organized into pairs for collaborative work.

Both lessons 4 and 5 provide an introduction to the management of string-like type of data and the classical workflow for the creation of models with ML algorithms, putting in practice what was previously learned in lesson 3. The idea is to (at least) familiarize with basics techniques related to development of supervised learning.

Although we presented, as first simple case-study, the SVM algorithm with a representation based on bag of words/TF-IDF weights, during lesson 5 we mentioned the current state of the art of the algorithms used to solve NLP tasks, and we encouraged the reflection on the best features that could help build a hate speech classifier.

Lesson 5 concludes with the administration of a final evaluation test (post-test), the analysis of which will be discussed in detail in Sections 5.1 and 6.1. This provides valuable feedback on the students' understanding and the effectiveness of the module.

*Lesson 5 in brief:* Colaboratory, Python, pandas, scikit-learn, CountVectorizer, TF-IDF, SVM, agreement/disagreement, accuracy, post-test.

At the very end of the whole laboratory, an anonymous survey questionnaire on satisfaction was administered (see a detailed analysis in Section 6.1).

## 5 Hands-on experience

The laboratory today[10] counts six editions, during which we adapted methodologies (Section 3) and activities (Section 4) to the different settings we encountered over the years, monitoring both students' and teaching strategies progresses.

[10]Time of writing: June 2024.

### 5.1 Evaluation

Since the first edition, at the end of the last lesson, we asked students to fill a survey questionnaire to express their overall satisfaction towards the laboratory and the degree of interest in the topics of the course. Students' feedback has been useful to map the adaptability of the methodologies to different settings, and what would need to be changed in order to make the lab more effective and appealing.

In addition, starting from the third edition, we built two tests to assess the degree of assimilation of the main concepts covered during the course, specifically a test of prior (pre-test) and final knowledge (post-test), to be administered respectively before the beginning of the laboratory, and at the end of the 10-hour cycle of lessons. The tests consisted of four kinds of questions:

**1. True/false:** evaluated as correct (1 point) or wrong (0 points).

**2. Multiple choice:** evaluated as right (2 points), partially right (1 point) or wrong (0 points).

**3. Questions that require fairly short answers:** evaluated as right (2 points), partially right (1 point) or wrong (0 points).

**4. Open questions that require a long answer:** evaluated on a scale ranging from 0 to 5.

Both the pre- and post-test were composed by questions related to different topics and categories of concepts dealt with during the lab, corresponding to the modules described in Section 4: i) *Sociology/Civics and Hate Speec*h (C); ii) *Computational Linguistics* (CL); iii) *Computer Science/Programming* (CS). Table 2 in Appendix A, provides examples for each of these categories, together with examples of the assigned notes for the open questions.

Most of the questions in the pre- and post-test overlapped in order to assess students' progress, together with the effectiveness of the laboratory. The pre-test was delivered before the introduction of ourselves and of the course (see Section 4.1), since we wanted to map their level of knowledge on the topics of the laboratory to actively engage the participants right away. The post-test was administered on the last lesson (see Section 4.5), or given by the last day as homework with a hard deadline (and help from the local teachers, for the deadline to be respected). It was presented to students as a proper assessment test, in order to encourage them to participate seriously and with

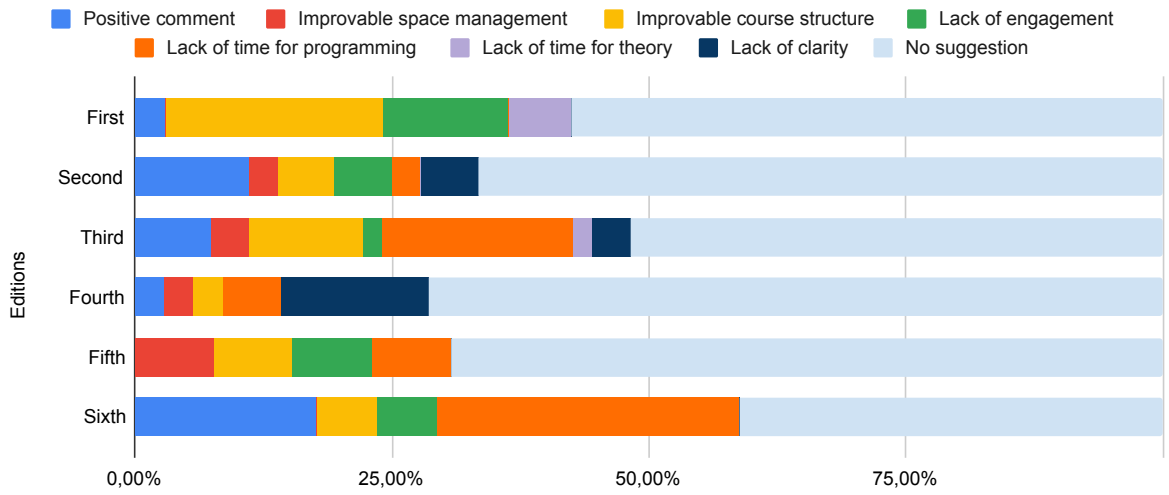


Figure 2: Grouped answers to the question: *Do you have any comments, suggestions, or constructive criticism that would be helpful in organizing future #DeactivHate laboratories?*

commitment. For the 3rd and 4th edition, both the questionnaires were held on the Moodle platform of our main affiliation (University of Turin). For the 5th and 6th edition they were held on Google Forms, since, post-2020, many schools began using Google Classroom as suite for online teaching.

## 6 Challenges and Lessons Learned

After six editions and seven teachers involved during the years, we want to share our considerations on challenges and lessons learned, since we believe it can be useful to open a deeper reflection on teaching NLP and offensive language detection nowadays in high schools. To carry out this analysis, we examined the answers to the anonymous survey **questionnaire on satisfaction**, and the results of the **pre- and post-tests**. Then we gathered together, sharing thoughts that emerged from reading the results, recollecting the experiences of each edition. All the questionnaires represented useful instruments to assess the effect of our methodologies in different settings, to summarize the challenges we were able of addressing during the years, and to highlight those that are still open.

### 6.1 Addressed challenges

We analyzed the anonymous opinions received from students in the survey questionnaire, and we grouped the replies in thematic groups. In Figure 2 we show the results.

In particular, we noticed a major difference related to time management between the online (first, third, and fifth) and the offline editions (second, fourth, and sixth). The laboratory started during the period of the COVID-19 pandemic outbreak, which forced us to deal with online teaching since the beginning (even though the laboratory was originally conceived to be held *in praesentia*). As teachers, we perceived a difference in the students' responsiveness in respect to offline teaching, specifically worsening time management.

Online teaching was particularly challenging in edition 1 because, in the first lessons, students were all connected from a single computer, making the interaction often filtered through the teacher in the classroom. The same happened in the fifth edition by necessity of the school, with the additional problem of having two classes of different levels merged sharing the same room, thus leading to the request for an improvement in space management (see Figure 2). These experiences taught us how a one-on-one interaction with students online is still preferable than having the whole class connected to one device, facilitating the possibility of engagement and helping them not to get lost, especially during the lectures dedicated to coding.

Looking at the pre- and post-test results in Figure 3 (administered from the third edition on, as referenced in Section 5), it is possible to observe an improvement in all the modules and editions. In the fifth edition, we noticed a higher percentage of students who have not assimilated concepts from the CS module. This result can also be associated with the fact that we worked with classes of two different levels at the same time, having students with different computer skills.

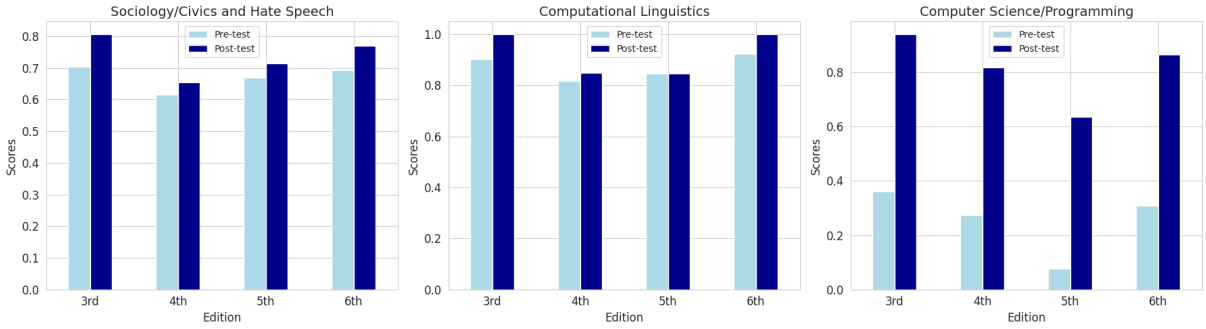


Figure 3: Mean of correct answers in pre- and post-tests for the subset of students who completed both tests (120).

Similarly, in the sixth edition there were students from different classes, since they could choose the course as *school-work experience*[11] on a voluntary basis, and the lab was an extracurricular activity for them. In this case, the improvements between the pre- and post-test were consistent. On the other hand, the strong request for more time for programming (Figure 2) could be linked to the fact that the teachers dedicated part of the fifth lesson to a visit to the buildings of the Computer Science Department of the University of Turin, thus 'sacrificing' time that would be typically dedicated to coding.

Despite these issues, the 6th edition showed us the positive aspect of having a class of people who volunteered to partake in the lab and, therefore, expressed an active interest on these topics, as demonstrated by the higher percentage of positive comments (Figure 2) and satisfaction.

A big challenge we encountered in the 4th edition was the presence of negative social bubbles in the classroom, and their influence in approaching hateful content, specifically linked to the figure of a well-known hate spreader and misogynist. To address this issue (also acknowledged by local teachers), we spent more time on lessons dedicated to the definition of hate speech and hateful content, significantly engaging with the class; thus, reducing the available time dedicated to CS and coding. This specific situation might be the cause of high scores in "lack of clarity" (refer to the dark blue portion in the graph, see Figure 2).

Moreover, we decided to share informative content on the topic with the local teachers, specifically Bold Voices advice[12] spread in Italy via the newspaper "Internazionale"[13].

[11]It is a compulsory activity foreseen in some types of higher education institutions in Italy, after one of the last Education reforms.

[12]https://www.boldvoices.co.uk/

[13]https://www.internazionale.it/

## 6.2 Open challenges

Throughout these years we addressed multiple challenges, nevertheless, there are still open issues that need to be discussed and worked out.

For instance, a major difficulty we found from the 4th edition on was to balance the introduction of NLP basics, and students curiosity towards more complex models such as LLMs, which are now part of their daily life. In the fifth edition, we dedicated around 10 minutes of the last lesson to introduce a visual article published by the Financial Times on the basics of generative AI[14], also adding the source to the advanced materials. This attempt was taken positively, but a more effective strategy to the entry of generative AI into the everyday lives of our students is definitely needed.

Considering the overall interest towards more hours of programming, another open challenge intends to better balance the second part of #DEACTIVHATE, introducing an additional (sixth) meeting, and working step by step by launching the programming part already from the second lesson, if the rooms and tools of the schools allow for it. We believe that delivering practical coding exercises in parallel with the theory lessons would lead to a more engaging setup. Furthermore, an extra lesson would allow us to delve into unsupervised learning, providing a comprehensive understanding of fundamental NLP concepts. It could also introduce alternative classification methods like multilayer perceptrons or transformer architectures such as BERT, offering at least a basic introduction to these (slightly more) advanced topics.

Another challenge, linked to the fact that most of the editions of the laboratory were part of a larger school guidance project, is to harmonize

notizie/anna-franchin/2023/04/07/andrew-tate-misoginia-violenza

[14]https://ig.ft.com/generative-ai/

this objective and keep it always updated with the involved students and teachers, reserving a proper time and space for it.

Finally, we are aware that students' awareness changes according to social and cultural factors, so it is important to make the laboratory flexible, and able to meet the needs and interests of each group we work with.

## 7 Conclusions

Our paper outlines the development and implementation of the #DEACTIVHATE laboratory, aimed at empowering high school students to address hate speech through computational thinking and NLP techniques. The laboratory's goals include introducing students to NLP techniques, raising awareness about ethical issues in the digital world, and fostering responsible technology usage. Through six editions of the laboratory, we have reached a diverse group of students, adapting methodologies and activities to different settings and backgrounds.

The related work section contextualizes our project within the broader academic landscape, highlighting the importance of automatic hate speech detection and educational initiatives for promoting responsible digital citizenship. Our approach incorporates practical exercises and utilizes platforms like Google Colaboratory to provide hands-on experience with NLP tools.

We describe in detail the teaching goals and methodologies employed in the laboratory, which include collaborative reading sessions, matrix design and analysis, practical coding exercises, and real-life scenario exploration on social media. Each lesson is designed to progressively build students' understanding of hate speech detection and NLP techniques.

The paper also presents the results of evaluations conducted throughout the editions, including pre- and post-tests administered to assess students' knowledge and the effectiveness of the laboratory. Challenges encountered during the implementation of the laboratory are discussed, along with lessons learned and open challenges for future iterations.

## Ethics Statement and Limitations

This paper has limitations, primarily stemming from our positionality as NLP academic researchers based in Northern Italy, which inherently introduces cultural and societal biases, as discussed in the first part of the paper. Secondly, it is crucial to consider that our theoretical framework concerning *Hate Speech* within the #DEACTIVHATE laboratory is situated within a European context. This framework refers to legislation and directives derived from the EU, as well as broader statements from the European Commission against Racism and Intolerance (ECRI).

• Different socio-cultural environments can influence the manifestation and perception of hate speech, as well as the effectiveness of various deactivation strategies. Therefore, while our insights contribute valuable knowledge, we recognize that they might (vastly) differ in contexts outside the one we operated in.

• The Wheel of Privilege was originally developed within the U.S., therefore it was adapted to our framework by, for instance, substituting *English* with *Italian* in the Language section of the Wheel of Privilege. We also noticed the absence of a 'slice' regarding Religion. We believe that other adjustments might be necessary, depending on the context in which this laboratory will be taught.

• The Wheel of Privilege was originally developed within the U.S., therefore it was adapted to our framework by, for instance, substituting *English* with *Italian* in the Language section of the Wheel of Privilege. We also noticed the absence of a 'slice" of the pie regarding Religion. We believe that other adjustments might be necessary, depending on the context in which this laboratory will be taught.

• We acknowledge that some activities might be triggering; therefore, we recommend careful consideration of the teachers. For instance, the activity carried out in Lesson 2 of researching hateful messages throughout social media pages, takes place after thorough reflection on the target and potential consequences.

• With our background and experience with this phenomenon, both as researchers and activists, we believe it is important it is crucial to highlight the problem rather than hide it. The issue of online hate is widespread, and young people are exposed to it daily, making it essential to address it with awareness and preparedness. Furthermore, the class should be designed to be a safe space for everyone, with precautionary measures in place and trigger warnings always provided (with the help of local high school teachers).

## Acknowledgements

The authors would like to thank the coordinators for their engagement in the start of the laboratory, for providing the first contacts with schools and for securing initial funding. Furthermore, the authors want to extend their thanks to all the high school professors that opened their doors and helped us deliver #DEACTIVHATE and monitor the students throughout the duration of the lab. Finally, a big thank to all the students who actively participated: without them this laboratory would not even exist.

The laboratory is supported by the *Commissione Orientamento e Informatica nelle Scuole* from the Computer Science Department of the University of Turin. The teaching activities are funded by the project *Piano Lauree Scientifiche* as part of the activities of the Computer Science Department of the University of Turin (MEOR_POT_PLS_23_01).

## References

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.

Alessandra Teresa Cignarella, Simona Frenda, Mirko Lai, Viviana Patti, and Cristina Bosco. 2023. #DeactivHate: An Educational Experience for Recognizing and Counteracting Online Hate Speech. *IJCoL. Italian Journal of Computational Linguistics*, 9(9-2).

Simona Frenda, Alessandra Teresa Cignarella, Marco Antonio Stranisci, Mirko Lai, Cristina Bosco, Viviana Patti, et al. 2021. Recognizing Hate with NLP: The Teaching Experience of the #DeactivHate Lab in Italian High Schools. In *Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, volume 3033, pages 1–7. CEUR-WS.org.

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 Computational Approaches to Social Modeling (ChASM 2014)*, June 23–26, 2014, Bloomington, IN, USA.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

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.

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.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th Conference on Language Resources and Evaluation (LREC2018), May 2018, Miyazaki, Japan*, pages 2798–2895.

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.

# A Appendix

| Edition | Mode | Period | Type | Grade and age | N. of students |
|---|---|---|---|---|---|
| 1st | online | April-June 2021 | humanities | III (15/16 y.o.)<br>IV (16/17 y.o.) | 21<br>14 |
| 2nd | in person | October-December 2021 | humanities | III$\alpha$ (15/16 y.o.)<br>III$\beta$ (15/16 y.o.) | 20<br>26 |
| 3rd | online | February-March 2022 | technical | III (15/16 y.o.)<br>IV (16/17 y.o.)<br>V (17/18 y.o.) | 25<br>20<br>19 |
| 4th | in person | April-May 2023 | technical | IV$\alpha$ (16/17 y.o.)<br>IV$\beta$ (16/17 y.o.) | 18<br>24 |
| 5th | online | January 2024 | technical | IV (16/17 y.o.)<br>V (17/18 y.o.) | 28 in total |
| 6th | in person | February 2024 | technical | III (15/16 y.o.)<br>IV (16/17 y.o.)<br>V (17/18 y.o.) | 17 in total |

Table 1: Details of the editions of the laboratory. In the second and fourth edition, we taught to two different classes of the same grade ($\alpha$ and $\beta$).

| **Question** | **Type** | **Topic** | **Example open answer** | **Vote** |
|---|---|---|---|---|
| By reading the following text you decide whether it contains hate speech (hs) or does not contain any (non-hs). | true/false | CL | | |
| How is text written in natural language processed by a machine/computer? Choose the alternative | multiple choice | CS | | |
| The following text contains at least one form of hate speech. Choose the discriminatory phenomenon you think best from the options below and explain why. [Racism, Misogyny, Sexism, Ageism, Homophobia, Abilism] | short answer | C | This is a form of ageism because generalizes on the age of the followers | 0 |
| *"How can you put up such a vulgar picture, shame on you, you are not up to being followed by children, you should not set such an example to an audience of kids/children following you"* | | | misogyny, physical appearance is judged and the content of the photo is deemed "vulgar" | 2 |
| A practical example of an algorithm in everyday life is... | long answer | CS | A practical example of an algorithm in everyday life is work. | 1 |
| | | | To fix my hair for example I do a series of "operations" that together define the algorithm: 1) I take the hair dryer; 2) I make sure my hair is completely dry; 3) I take the foam; 4) I spread it on my hair so that it is a bit curly; 5) I take the hair dryer again; 6) I blow dry my hair; 7) I take the gel; 8) I spread it on my hair and fix it calmly hair by hair; 9) I put down the gel, the foam and the hair dryer. | 5 |

Table 2: Example of different types of questions in respect to the three main topics of the course.

## B Available Materials

All the materials created for the #DEACTIVHATE laboratory are available at the following link: https://github.com/deactivhate. Below, we provide a complete list of the files contained in the GitHub repository. First, a general document explaining "how we structured the course", and then 5 folders, one per lesson, containing the following materials:

**Lesson 1:**
• Icebreaker JamBoard
• Introduction to #DeactivHate (slides)
• Pre-test

**Lesson 2:**
• Social and personal identity + pyramid + hate speech definition (slides)
• Forms of hatred (slides)
• Tweets containing Hate Speech (spreadsheet)

**Lesson 3:**
• Machine Learning workflow - 1st part (slides)[15]
• Machine Learning workflow - 2nd part (slides)
• Bag of words matrix (spreadsheet)

**Lesson 4:**
• Colab + Python (slides)
• Introduction Colab Python (interactive python notebook)*

**Lesson 5:**
• Supervised Classification (interactive Python notebook)*[16]
• Extra material on Machine Learning workflow
• Post-test

* The interactive notebook files for coding contain cells of code with one or more possible solutions of the task. With the purpose of introducing students to manage strings and creation of NLP models, during the lessons we used a version of these files without solutions provided.

**> > >** The ideal instructor(s) for teaching this course should have at least an expertise in the following topics: hate speech detection and legislation, basics of natural language processing, high school teaching.

[15] In slide 25 of this presentation, we mention the already obsolete Twitter API as possible software to collect data online. Probably best if updated.

[16] The dataset used inside this interactive notebook contains Italian texts. Datasets in other languages and on different topics can be found, for instance here: https://live.european-language-grid.eu/.