IJCoL
Italian Journal of Computational Linguistics

6-1 | 2020
Emerging Topics at the Sixth Italian Conference on Computational Linguistics

“Contro L’Odio”: A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media

Arthur T. E. Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, Giovanni Semeraro and Marco Stranisci

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/ijcol/659
DOI: 10.4000/ijcol.659
ISSN: 2499-4553

Publisher
Accademia University Press

Printed version
Number of pages: 77-97

Electronic reference
Arthur T. E. Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, Giovanni Semeraro and Marco Stranisci, ““Contro L’Odio”: A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media”, *IJCoL* [Online], 6-1 | 2020, Online since 01 June 2020, connection on 24 December 2021. URL: http://journals.openedition.org/ijcol/659 ; DOI: https://doi.org/10.4000/ijcol.659

"Contro L'Odio": A Platform for Detecting, Monitoring and Visualizing Hate Speech against Immigrants in Italian Social Media

Arthur T. E. Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo*
Università degli Studi di Torino

Cataldo Musto, Marco Polignano, Giovanni Semeraro**
Università degli Studi di Bari "Aldo Moro"

Marco Stranisci†
Acmos
Università degli Studi di Torino

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 have been developed which are centered on the platform and focused on the deconstruction of negative stereotypes against immigrants, Rom and religious minorities, and on the creation of positive narratives. The data collected and analyzed by the platform are also currently used for benchmarking activities within an evaluation campaign, and for paving the way to new projects against hate.

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 much attention in the NLP community, but also in public institutions and private companies. As a privileged place for the expression of opinions, feelings and emotions, social media are particularly suitable for conveying not only generic expressions of offensiveness and

* Dept. of Computer Science - C.so Svizzera 185, 10149, Turin, Italy.
E-mail: name.surname@unito.it

** Dept. of Computer Science - Via E. Orabona 4, 70017, Bari, Italy.
E-mail: name.surname@uniba.it

† Dept. of Computer Science - C.so Svizzera 185, 10149, Turin, Italy.
E-mail: marcoantonio.stranisci@unito.it

hatred, but hate speech, which affects individuals or groups of people because of their belonging to a vulnerable category, that is as they are characterized by a particular race or ethnicity, religion, gender or sexual orientation. In the background there are the tensions manifested at social level in relation to events and situations, that act as triggers of hate speech, sometimes giving rise to harmful hate campaigns (Florio et al. 2020). Considering that social media allows for a wide and rapid dissemination of messages, the extreme expressions of verbal violence and their proliferation on the network are gradually taking shape as mandatory social emergencies. As such, they should be addressed through coordinated interventions between institutions within individual states or at the level of larger communities such as European Union[1].

There are several problems connected with the delicate task of detecting HS: 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 (thus restricting the right of freedom of opinion) (Pamungkas, Basile, and Patti 2020), 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"[2] 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 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[3] and Vox[4], and two academic research groups, respectively from the University of Bari and Turin.

This paper focuses on the technological core of the project and on its impact on educational and research activities. The "Contro l'Odio" Web platform combines computational linguistics analysis with visualization techniques, in order to provide users with an interactive interface for exploring the dynamics of hate speech against immigrants in Italian social media. Three typical targets of discrimination related to this topical focus are taken into account, namely migrants, Muslims and Rom, since they exemplify discrimination based on nationality, religious beliefs and ethnicity, respectively. Since October 2018 the platform analyzes daily Twitter posts and exploits temporal and geo-spatial information related to messages in order to ease the summarization of the hate detection outcome. The platform has also been used by the civil society organization partners for educational purposes in courses for high school students, where the monitoring functionalities enabled by the platform supported the work of educators with the final aim of deconstructing negative stereotypes against immigrants, Rom and religious minorities, and creating positive narratives.

1 See for instance the Code of Conduct on countering illegal hate speech online issued by EU commission (EU Commission 2016).
2 https://controlodio.it/
3 http://acmos.net/
4 http://www.voxdiritti.it/

The paper is organized as follows. The next section surveys the main contributions in the field. Section 3 presents the architecture of the "Contro l'Odio" monitoring platform, illustrates the data collection process, the hate speech detection engine and presents two kind of analysis: the first one devoted to shed some light on the topics of discussion emerging from the data collected, the second one aimed to analyze the social network and detect users acting as haters in the online debate considered. Section 4 focuses on the data visualization tools implemented, including the interactive hate maps. Section 5 summarizes the educational activities carried out in high schools and centered on the platform, while the last section includes some conclusive remarks on the work done and on the immediate and future impacts in different directions.

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 (Schmidt and Wiegand 2017; Fortuna and Nunes 2018) and a systematic review were recently published on this topic (Poletto et al. 2020). For what concerns Italian, a few resources have been recently developed using data from Twitter (Sanguinetti et al. 2018; Poletto et al. 2017; Comandini and Patti 2019), Facebook (Del Vigna et al. 2017) and Instagam (Corazza et al. 2019). A multilingual lexicon of hate words has also been developed (Bassignana, Basile, and Patti 2018), called HurtLex[5]. 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. These include in particular HatEval at SemEval 2019 (Basile et al. 2019), AMI at IberEval 2018 (Fersini, Rosso, and Anzovino 2018), HaSpeeDe and AMI at EVALITA 2018 (Bosco et al. 2018; Fersini, Nozza, and Rosso 2018), with their follow up proposed at EVALITA 2020[6] (Fersini, Nozza, and Rosso 2020; Sanguinetti et al. 2020).

For a more complete overview of the available HS resources, including lexica and benchmark datasets, in Italian and in other languages, we refer to Poletto et al. (2020).

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. 2020) has been exploited to provide a fine-grained classification of the nature of the hate speech 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.

5 http://hatespeech.di.unito.it/resources.html
6 http://www.evalita.it/2020

Monitoring and countering HS is a shared goal with several recent European and Italian projects, which have focused on different hate targets, different languages, countries and territories, differentiating themselves for the granularity of the detection, time frame taken into consideration, the possibility of offering daily monitoring or *ex post* analysis, and, finally, as regards the visualization techniques provided to inspect the results of the monitoring and quantify the phenomenon, with static or dynamic and interactive maps. Among those, the *CREEP* project[7] on monitoring cyberbullying online (Menini et al. 2019), with an impact also on the Italian territory, *HateMeter*[8], with a special focus on Anti-Muslim hatred online and on opposing hate content with counter-narratives (Chung et al. 2019), the MANDOLA project[9] providing an infrastructure enabling the reporting of illegal hate-related speech (Paschalides et al. 2020), the *Hatred Barometer*, coordinated by the Italian section of Amnesty International[10], and the *Geography of Hate* project[11] in the US.

Considering the Italian social media discourse on immigration, it is worth mentioning a related study where the linguistic analysis of Twitter data is combined with the social network analysis of the debate about immigration (Vilella et al. 2020). The work shows that communities tend to display segregation and that the most frequently occurring bi-grams found in texts within each community can be used as signal to understand their stance towards migrants.

It is worth pointing out that the implementation of effective monitoring tools cannot disregard the problem of defining what hate speech is. This issue, as already emerged in Poletto et al. (2020), is strongly posed as a question to be addressed. Finding a univocal and satisfactory definition of the hate speech phenomenon is challenging, and one of the main difficulties lies in drawing boundaries between hate speech and other broader phenomena such as offensive language. A starting point for the operational definition of hate speech that guided the work within the "Contro l'Odio" project is the definition proposed 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, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin"*. According to the operational definition that we have chosen as a reference, in order to speak of hate speech in a message, the presence of two elements is essential: a) the post must be addressed to a target, individuals or entire groups belonging to a vulnerable category; b) the post must intentionally incite / spread / promote hatred towards that category. This means that, to recognize the presence of HS, it is essential that hate speech is strictly connected with the idea of harm, discrimination or true violence against a target belonging to a vulnerable category, where the identification and variation over time of vulnerable or most at risk targets in the Italian context can be an interesting output of monitoring, valuable for legislators and policy makers.

7 http://creep-project.eu/
8 http://hatemeter.eu/
9 http://mandola-project.eu/
10 Barometro dell'odio: https://www.amnesty.it/barometro-odio/
11 http://www.antiatlas.net/geography-of-hate-geotagged-hateful-tweets-in-the-united-states-en/

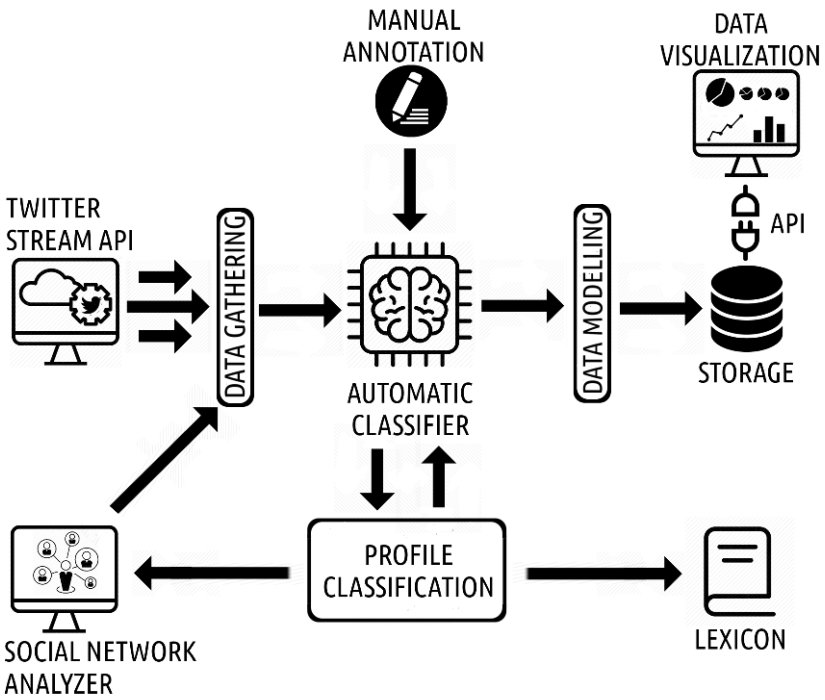

Figure 1
Architecture of the 'Contro l'Odio' platform

3. The Contro l'Odio Monitoring Platform

The architecture consists of four main modules, shown in Figure 1. The data collection module gathers the tweets by using the Stream Twitter API and filters them by keywords. Next, the automatic classifier module automatically annotates the presence of HS in the filtered tweets, relying on a supervised approach. This information is then exploited by the profile classifier module, whose goal is to identify, given a lexicon, *haters* based on the content they posted on social media. This process can be further iterated by also gathering content posted by other users in the social network. Finally, the storage module acquires the annotated tweets aggregating them by time and place in a database. The last module, implemented by relying on a *node.js* server, exposes the APIs that are requested by the front end.

3.1 Data Collection

We started collecting tweets from October 1st 2018 by using the Twitter's Stream API. The streaming is filtered using the vowels as keywords and the alpha-2 code *it* as language filter. About $800,000$ 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, Rom and religious

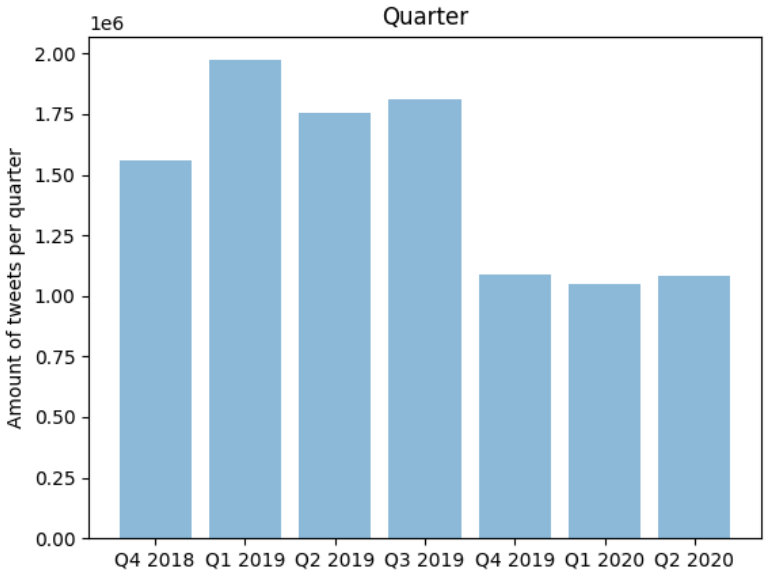

Figure 2
The amount of tweets addressed to vulnerable categories from October 1st 2018 to June 30th 2020

minorities. More precisely, from October 1st 2018 to June 30th 2020, 11,431,792 tweets addressed to these categories have been collected. Figure 2 shows the distribution of tweets per quarter.

Despite its effectiveness in showing a trend about a relevant topic in Italy, such as the debate about minorities, an arbitrary keyword selection entails two issues. First, a number of out of topic tweets is retrieved due to the ambiguity of words like *terrorista* (terrorist), which can be referred to non-religious phenomena, *straniero* (foreigner), that could pertain to finance or tourism, and *nomadi* (nomads), which is also the name of an Italian pop band. Considering manual annotation rounds organized by the Hate Speech Monitoring Group from 2018 to 2020, the 4.5% of tweets have been labeled as out of topic by human judges. Second, the choice of such general keywords for data gathering risks not to be effective in monitoring the evolution of the debate around vulnerable minorities. However, although the drawbacks of keyword-based approaches are known to researchers (Davidson, Bhattacharya, and Weber 2019), there are currently no clear alternatives to this technique, as discussed in Poletto et al. (2020).

3.2 The Hate Detection Engine

In order to automatically label the tweets, we developed a supervised binary classifier to predict the presence of HS in text. The final version of our hate detection engine has been obtained by performing different experimental runs. In each of them, we decided to vary the model, the pre-processing of training data, or the size of the original dataset, expanding it with new data. Below, we first provide an overview of the original dataset used in our experiments; we then describe more in detail the system characteristics and main results.

Training data. The dataset used to train the classifier is the one described also in Florio et al. (2019), which, in turn, consists of the Italian Hate Speech Corpus (Poletto et al. 2017; Sanguinetti et al. 2018), referred as IHSC henceforth, with the addition of brand new data collected from TWITA (Basile, Lai, and Sanguinetti 2018) and dating back to 2017. The dataset mentioned in Florio et al. (2019) consists of more than 15,000 tweets, but after a further manual cleaning we retained a final version consisting of approximately 11,000 tweets. The collection comprises tweets retrieved with the keyword-based method sketched above, resorting to a hand-crafted list of neutral keywords referred to three minority groups in particular, who were deemed as typical HS targets in the Italian context: immigrants, Muslims and Rom.
As regards the annotation process, while the IHSC underwent a mixed procedure that relied both on experts and crowdworkers, the additional data introduced in Florio et al. (2019) was completely annotated by Figure Eight (now Appen[12]) contributors. In both cases, the same annotation scheme and guidelines[13] were followed to build the dataset, so as to allow a greater consistency in the annotation choices. The scheme in particular has been conceived so as to provide a proper representation of hate-related phenomena, also with the aim to explore the actual correlation among them and HS. For this purpose, other dimensions were included in the annotation scheme, such as aggressiveness, offensiveness, stereotype and irony, the latter being considered in this context as a potential linguistic device used to convey, or rather mitigate hateful content.

System description and setups. The "Contro l'Odio" pipeline currently employs a Support Vector Machine (SVM) classifier with one-hot unigram representation as feature vector, trained on the Italian Hate Speech dataset mentioned above. We evaluated the model performing a 5-fold cross-validation experiments on such corpus obtaining **0.81** (0.70 for the class *hate speech*) precision and **0.81** (0.67 for the class *hate speech*) recall ($F_{avg} = 0.80 \pm 0.01$).

Due to the fact that the main goal of "Contro l'odio" is developing a sort of *Observatory* to monitor hate speech in Italy over the time, we focused on further experiments in order to evaluate different models using additional test sets in a diachronic perspective. For this purpose, we specifically created 6 new test datasets composed of 2,000 tweets, one for each month, from September 2018 to February 2019. In the first experimental run, we decided to evaluate the ability of some state-of-the-art classification models to correctly classify HS on our data. We thus assessed their performance using the first one of the six test sets mentioned above, i.e. the one comprising tweets from September 2018. We considered as a primary evaluation metric the F1 score achieved for the Hate Speech class. The choice of F1 on HS class only as a metric for comparison among models is supported by the idea that for a daily use scenario of the Hate Detection Engine, we are more interested in a system able to correctly detect HS rather than non-hateful content. During the experimental run we varied the model used for the classification task, using the following techniques: Random Forest, Decision Tree, Support Vector Machine (SVM), Long Short Term Memory (LSTM), Convolutional Neural Network (CNN), a CNN layer followed by an LSTM, BERT trained on the Italian language (AlBERTo) (Polignano et al. 2019b, 2019a). These models are considered as the state of the art for text classification tasks in many different scenarios, including sentiment

12 https://appen.com
13 Also available here:
https://github.com/msang/hate-speech-corpus/blob/master/GUIDELINES.pdf

Table 1
Results of the first experimental run obtained by varying the classification model. Test data: September 2018.

Model	F-score_HS_yes	F-score_HS_no	F-avg score	Precision HS yes	Recall HS yes
SVM	**0.510**	**0.793**	**0.652**	**0.566**	**0.464**
Random Forest	0.024	0.803	0.413	0.348	0.012
Decision Tree	0.321	0.774	0.545	0.446	0.249
LSTM	0.153	0.768	0.46	0.308	0.675
CNN+LSTM	*0.503*	*0.724*	*0.6335*	*0.496*	*0.511*
AlBERTo	*0.583*	*0.714*	*0.648*	*0.487*	*0.724*

analysis, entailment, aspect term extraction and many more. The results obtained in this first experimental run are reported in Table 1.

For ease of comparison, in this phase, we did not focus on the best strategy to encode the text. Therefore, we decided to keep the text representation as a fixed variable. Specifically, we used unigrams and bigrams identified in each tweet as its representation for SVM, Decision Tree, and Random Forest algorithms. For the algorithms of LSTM and CNN+LSTM, we encoded the text using a pre-trained word embedding space. In particular, we used the word2vec embedding space learned on 3 billion of news collected by Google News [14]. By observing the results in Table 1, it is possible to note that the SVM model performs better than the others in terms of macro-average F1 score. On the other hand, considering the F1 score for the HS class, SVM is the second best performing model after AlBERTo. The difference in results between these two models is minimal and it is around 0.073 only. Some results close to those of SVM and BERT for the F1 score for class HS are also obtained from the model based on CNN and LSTM. However, the performances are very low compared to the result obtained in using a cross-validation on the training set.

Considering the SVM model obtained the highest result, we decided to evaluate different configurations by varying the strategy of text pre-processing. In particular, we reduced the number of n-grams to 1, and we evaluated the following settings: one configuration that uses a one-hot unigram representation (SVM 1 grams); one that uses the TF-IDF score (SVM 1 grams tf-idf); one based on the one-hot representation of unigrams concatenated with some pre-processing statistics and the total number of tokens found into the tweet (SVM 1 grams pre-processing length); one that concatenates, at the previous representation, the lexicon of Hurtlex as an additional one-hot vector (SVM 1 grams hurtlex pre-processing length). Moreover, we also evaluated the SVM configuration that uses uni and bi-gram one-hot encoding strategy concatenated with the pre-processing statistics and the total number of tokens found into the tweet (SVM 1-2 grams pre-processing length). The pre-processing phase is performed by removing numbers, hashtags, links, mentions, punctuation marks, and lowercasing the text. The statistics about the presence of those lexical elements, and their amount are fundamental parts of the text encoding strategy here evaluated. In particular we considered as pre-processing statistics the following fields: num-

14 https://github.com/mmihaltz/word2vec-GoogleNews-vectors

Table 2
Results of the second experimental run obtained by varying the text encoding for the SVM. Test data: September 2018.

Model	F-score_HS_yes	F-score_HS_no	F-avg score	Precision HS yes	Recall HS yes
SVM 1 grams	*0.55*	*0.795*	*0.672*	*0.572*	*0.529*
SVM 1 grams tf-idf	0.467	0.802	0.634	0.582	0.390
SVM 1-2 grams pre-processing length	0.550	0.795	0.671	0.572	0.529
SVM 1 grams pre-processing length	0.467	0.802	0.638	0.582	0.390
SVM char 2-5 grams	0.525	0.781	0.655	0.541	0.511
SVM 1 grams hurtlex preprocessing length	**0.577**	**0.722**	**0.652**	**0.485**	**0.696**

ber_of_hashtags, number_of_mentions, number_of_links, number_of_numbers, number_of_puntactionmarks, number_of_uppercase_letters.
As mentioned above, a multilingual lexicon of hate words was also used in these experiments, i.e. HurtLex. Starting from the Italian lexicon "Le Parole per Ferire" by Tullio de Mauro, the authors developed a computational lexicon and semi-automatically translated it into more than 50 languages [15]. The lexicon of this vocabulary was added as extra tokens of our vocabulary used for the textual representation of tweets.

Table 2 shows the results obtained by SVM based on the different text encoding strategies. We can observe that the model using only the one-hot vector of unigrams obtained the best results in terms of macro-F1. Nevertheless, taking into account the F1 for the HS class (more relevant in this context), we can observe an increase of performances of 0.027 for the model that also includes the pre-processing statistics, the length of the tweet, and the Hurtlex lexicon.

However, even though we experimented different configurations, the results obtained are low if compared with the state of the art. We hypothesize that the drop in performances is caused by the fact that the training set and the test set were created using tweets gathered in two distance time windows.

Therefore, as a final evaluation, we decided to investigate the robustness of the best models we obtained until this step for the HS class, i.e. *SVM 1 grams Hurtlex preprocessing length* and *AlBERTo* extending the training set with additional data. We performed five incremental steps for each model by increasing the training set with new tweets collected and annotated from September 2018 to February 2019. In particular, starting from the data collected from September 2018, we added these tweets to the training set and we evaluated the model on the following month, i.e. October 2018. The same strategy was used for the other months, such that each month we increased the training set available from the previous one by 2000 tweets.

The results presented in Table 3 and Table 4 show a common behavior between the two classification approaches. In detail, both show the decreasing performances whenever the learning of the model is performed in an incremental way. Overall, the SVM proves to be more stable than AlBERTo, obtaining F1 values for the HS class on average higher than those obtained by the latter. The cause of this behavior is attributable to the

15 http://hatespeech.di.unito.it/resources.html

Table 3
Results of the third experimental run obtained by evaluating AlBERTo on an incremental training setting.

Model	Training Set	Test Set	F-score_HS_yes	F-score_HS_no	F-avg score	Precision HS yes	Recall HS yes
AlBERTo	IHSC + 09/2018	10/2018	0.431	0.737	0.685	0.301	0.824
-	IHSC + 09/2018 + 10/2018	11/2018	0.402	0.714	0.662	0.268	0.809
-	IHSC + 09/2018 + 10/2018 + 11/2018	12/2018	0.406	0.711	0.662	0.272	0.820
-	IHSC + 09/2018 + 10/2018 + 11/2018 + 12/2018	01/2019	0.288	0.719	0.471	0.177	0.818
-	IHSC + 09/2018 + 10/2018 + 11/2018 + 12/2018 + 01/2019	02/2019	0.254	0.707	0.464	0.157	0.836

Table 4
Results of the third experimental run obtained by evaluating our best configuration of the SVM model on an incremental training setting.

Model	Training Set	Test Set	F-score_HS_yes	F-score_HS_no	F-avg score	Precision HS yes	Recall HS yes
SVM	IHSC + 09/2018	**10/2018**	0.445	0.752	0.624	0.330	0.681
-	IHSC + 09/2018 + 10/2018	11/2018	0.413	0.782	0.626	0.326	0.564
-	IHSC + 09/2018 + 10/2018 + 11/2018	12/2018	0.430	0.776	0.646	0.363	0.528
-	IHSC + 09/2018 + 10/2018 + 11/2018 + 12/2018	01/2019	0.367	0.732	0.531	0.282	0.525
-	IHSC + 09/2018 + 10/2018 + 11/2018 + 12/2018 + 01/2019	02/2019	0.335	0.729	0.515	0.245	0.529

nature of the data that adapts quickly to events that occur in a given period of time. For example, if events concerning landings of illegal immigrants happened, we could easily find an increasing number of hate tweets compared to this category, probably with a completely new vocabulary and never found in the previous months' tweets. A deeper analysis providing some experimental evidence of such hypothesis can be found in Florio et al. (2020).

The results of our experimental runs supported our initial idea to employ an SVM classifier, with one-hot unigram representation of tweets, as the core algorithm of our Hate Detection Engine.
The following are two examples of tweets correctly annotated by the Hate Detection Engine:

(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) Le vittime sono tutte uguali Cristiane, ebree, musulmane, atee. Siamo parte della stessa umanità!

The victims are all equal Christians, Jewish, Muslims, atheists. We are part of the same human race!

In example 1, the target is "religious minorities" and the author spreads and incites violence against Muslims. The tweet is thus classified as hateful. In example 2, the target is instead the "crimes against religious minority". In this case the tweet is promoting the equality among different religions and, consequently, the hate conditions are not detected.

3.3 Hate Speech Analysis

In order to gain a better understanding of the interacting phenomena in the corpus collected by the project, we performed a qualitative analysis supported by a rigorous statistical methodology. The aim of this study is to discover and analyze the topics of discussion emerging from the data, and their diachronic behavior. Our main statistical tool is the *polarized weirdness index* (Florio et al. 2020). This word-level measure is based on the weirdness index (Ahmad, Gillam, and Tostevin 1999), an intuitive and flexible technique which can be applied to several domain of knowledge, and text types. This automatic metric retrieves the more frequent and characterizing words within a given corpus (e.g., a repertoire of specialized language documents or a collection of texts that refers to a particular domain) in a particular time span. The idea behind this method is straightforward: the specific corpus is evaluated against a more general, and wider dataset. First, the relative frequency of each word in both collections is calculated. Then, the ratio between the two frequencies is computed. As a result, only the words that are frequent in the specialized corpus but not in the general one are ranked with the highest score. The weirdness index behaves similarly to metrics from Information Theory, such as Information Content (Pedersen 2010), and measures of frequency distribution similarity, such as Kullback-Leibler divergence (Kullback and Leibler 1951).

The polarized version of the weirdness index implements the same principle, but applied to annotated data. In particular, we treat the portion of the corpus annotated with $HS = 1$ as the specialized corpus, so to make the keywords relevant to the hate speech phenomenon emerge. We further distinguish between *Weak* Polarized Weirdness (WPW), where the comparison is made against a general corpus, from *Strong* Polarized Weirdness (SPW), where the contrast corpus is the complement of the specialized corpus in terms of annotation. We ran an experiment computing the weak polarized weirdness index on the full corpus from the project with the labels produced automatically by our classifier, and using as general corpus TWITA, the large-scale collection of Italian tweets (Basile, Lai, and Sanguinetti 2018).

Table 5 shows the results of our analysis. Each column represents a 3-month period, and contains the words with the highest WPW score from all the words in the corpus with absolute frequency greater than 15. We observe two phenomena worth investigating further. The first is a series of words that are not included among the original keywords but are present at the top of the list in each period of time. Among these, we notice several words related to the legal and administrative status of the conditions of immigrant, with a clear negative bias: *clandestini* (clandestine), *irregolari* (irregulars), *rimpatriare* (repatriate).

Secondly, there are tokens that appear only in specific time frames, suggesting a link with events relevant to the political debate. E.g., the word *globalcompact* has the highest polarized weirdness index during the first quarter of 2019, due to the UN pact about immigration. Another example is the verb *lucrare* (to profit, to speculate), highly

Table 5
Analysis based on the weak polarized weirdness index.

Q4 2018	Q1 2019	Q2 2019	Q3 2019	Q4 2019	Q1 2020	Q2 2020
clandestin*	globalcompact	irregolar*	impediremo	irregolari	mega-hub	finanzio
irregolar*	irregolari	#clandestin*	giorno	#clandestini	bomba-africa	irregolari
rimpatriare	#clandestin*	rimpatriare	#barconi	disgraziate	irregolari	#clandestini
#profughi	lucrare	rispedisce	irregolari	psicotici	#sbarchi	#profughi
politico-mafioso	#sbarchi	lucrare	#italianversion	clericale	rimpatriare	#sbarchi
lucrare	rimpatriare	#sbarchi	#sbarchi	#sbarchi	#clandestini	rimpatriare
#immigrati	incontrollati	clandestinamente	#clandestini	#profughi	#terroristi	#immigrati
barconi	#profughi	tagliagole	#profughi	#invasione	#invasione	gommoni
fraudolento	astenuta	sbarcare	#rockcover	rimpatriare	#immigrati	barchini
#rifugiati	sorvegliati	sbarchi	#immigration	clandestinamente	gommoni	rispediti
#accoglienza	#immigrati	#accoglienza	#refugees	#immigrati	barchini	barconi
astronave	barconi	gommoni	rimpatriare	sbarcheranno	sbarcheranno	intrecci
gommoni	clandestinamente	#immigrati	lucrano	#rifugiati	barconi	espelle
cpr	gommoni	barchini	#invasione	barconi	#accoglienza	accoglienza
#migranti	barchette	barconi	#toto	respingimenti	lucrano	#migranti

relevant in our corpus up to September 2019, when the country underwent significant changes in the political organization (change of government) and consequently political discourse.

We finally note that we calculate the polarized weirdness technique employing an automatically annotated corpus, in contrast with previous work (Florio et al. 2019). As a byproduct of this experiment, we consider the consistency of the most recurring tokens with the main topic as a sort of validation of the corpus creation methodology.

3.4 Social Network Analysis

Another outcome of the activities of the "Contro L'Odio" research project is HATECHECKER, a tool for the automatic detection of *hater* users in online social networks that exploits sentiment analysis and natural language processing techniques.

The hallmark of the tool is the focus on the detection of *hater users* rather than *hate speech*, as most of the current literature does. Indeed, HATECHECKER aims to analyze the users *as a whole* in order to identify those that usually spread and post hateful contents. Of course, the task of detecting hater users is clearly connected to the task of detecting hate speech. To this end, this component tackles the task by obviously exploiting the algorithms presented in the previous section.

In a nutshell, our tool implements a methodology based on three steps: *(i)* all the tweets posted by a target user are gathered and processed; *(ii)* sentiment analysis techniques are exploited to automatically label intolerant tweets as *hate speech*; *(iii)* a lexicon is used to classify hate speech against a set of specific categories that can describe the target user (e.g. racist, homophobic, anti-semitic, etc.).

Finally, the output of the tool, that is to say, a set of labels describing (if any of) the intolerant traits of the target user, are shown through an interactive user interface.

The whole pipeline implemented in the HATECHECKER tool needs some *textual content* posted by the target user to label the user as a *hater* or not. In the absence of textual content, it is not possible to provide such a classification. To this end, we exploited the Data Gathering component (see Figure 1) to collect the tweets posted by the user we want to analyze.

Next, we used sentiment analysis techniques to process all the content we previously extracted and to detect hate speech. However, in such a specific setting, the simple

exploitation of sentiment analysis techniques that provide a *rough* binary classification of the single tweets (*conveying/not conveying hate*) is not enough.

Indeed, HATECHECKER needs to answer two fundamental questions: *(i)* How can we label the user *as hater* or *non-hater* based on the tweets they posted?; *(ii)* How can we return a more fine-grained classification of the user (e.g. racist, homofobe, etc.) based on tweet they posted?

Both these issues are tackled by the PROFILE CLASSIFIER module (see Figure 1). As for the first question, a very simple strategy based on *thresholding* is implemented. In particular, we defined a parameter ϵ, and whenever the user posted a number of tweets labeled as *hate speech* higher than ϵ, the same user is labeled as an *hater*. Of course, several values for the parameter ϵ can be taken into account to run the tool, and we leave for future work the automatic tuning of such a parameter.

As for the second question, we used a *lexicon-based* approach to provide a fine-grained classification of users' profiles. Our methodology is based on the intuition that each category is described by a specific lexicon, and whenever a tweet posted by the user contains one of the terms in the lexicon, the user is labeled with the name of the category.

Formally, let $C = \{c_1, c_2 \dots c_n\}$ be the set of the categories (e.g. racism, homophobia, sexism, etc.) and let $V_{C_i} = \{t_1, t_2 \dots t_m\}$ be the vocabulary of the category C_i. Given a tweet T written by a user u, if one of the terms in V_{C_i} is contained in T, the user u is labeled with the category C_i. Even though we are aware that more sophisticated techniques exist, as for this first prototype we chose a strategy that provided us with a good compromise between effectiveness and simplicity. As future work, we will also consider the adoption of other mechanisms to identify *hater* users (*e.g.*, based on vector space representations and similarity measures).

To define the lexicon for each category, we relied on the research results of the Italian Hate Map (Lingiardi et al. 2020). In particular, we exploited the categories as well as the lexicon used in the Italian Hate Map Project, which consists of 6 different categories (*racism, homophobia, islamophobia, xenophobia, anti-semitism, sexism, abuse against people with disabilities*) and 76 different terms in total.

In order to (hopefully) enrich and improve the lexicon used in the Italian Hate Map project, we exploited HurtLex, the multilingual lexicon also used in our classification experiments (see Sect. 3.2). Specifically, we manually selected a subset of relevant terms among those contained in HurtLex and we merged the new terms with those contained in the original lexicon. In total, the complete lexicon contained 100 terms, 76 coming from the original Italian Hate Map lexicon and 24 gathered from HurtLex.

At the end of the previous step, the target user is labeled with a set of categories describing the *facets* of their intolerant behavior.

However, one of the goals of the project was also to investigate the role and the impact of the social network of the users in the dynamics of online *haters*. Accordingly, the SOCIAL NETWORK PROCESSOR gathers the entire social network of the target user and runs again (in background) the whole pipeline on all the *following* and *followers* of the target user, in order to detect whether other people in the social network of the target user can be labeled as *haters* as well. The goal of this step is to further enhance the comprehension of network dynamics and to understand whether online *haters* tend to follow and be followed by *other haters*.

4. Visualizing and Interacting with Estimated Hate

We have created a data visualization platform to support the analysis of the spread of hate speech on Twitter in Italy. This web-based platform has been designed for supporting interactive access to analyzed data and visualization of *hate maps*, a powerful tool that can be exploited in a variety of decisions humans must take everyday with respect to their behavior towards the community or other people, i.e. for democratizing the knowledge about immigration. The platform provides indeed a visual, easy-to-read representation of the analysis provided by the automatic detection engine, and it can support the different end users to gather the intelligence required, for example, to make informed decisions on local policies or support a continuous monitoring on online hate speech to prevent escalation of hate and violence that would potentially affect the stability of local communities.

Since we had to deal with many different types of data, we have developed multiple visualizations, each one designed to best represent the data domain.

4.1 Interactive Hate Maps

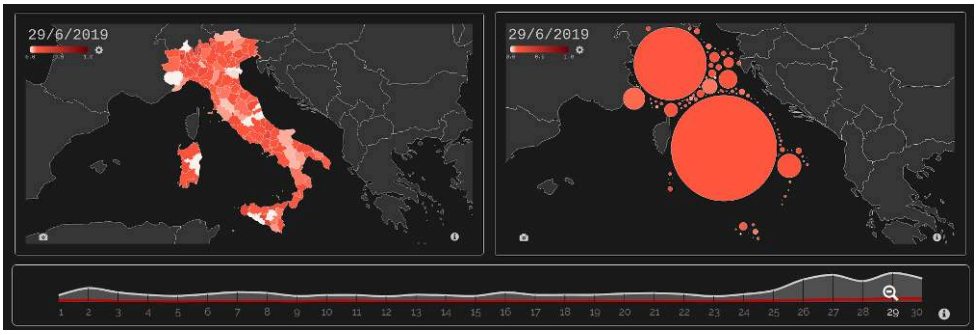

Figure 3
On the left, a choropleth map, on the right a Dorling cartogram. Both maps show, for each region, the spreading of HS in tweets on June 29, 2019. In the Dorling cartogram, the total number of tweets is represented by the size of each circle (this further information cannot be represented in a choropleth map).

The main view of the dashboard is a choropleth map and allows the user to explore the spatial dimension (regional and provincial level) of the dataset. *Choropleth maps* are often used to visualize the spatial distribution of aggregated data collected. The geographic distribution can be represented also through a *Dorling cartogram*, a technique for representing data for areas that eschews geography in preference for a geometric shape that represents the unit areas. A Dorling cartogram maintains neither shape, topology or object centroids and is an abstract representation of the spatial pattern of the phenomena being mapped. In figure 3 there is an example of how the choropleth map and the Dorling cartogram appear on June 29, 2019 when the migrant and NGO themes, in a single day, become viral in the public debate[16]. Color codes of both choropleth

16 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

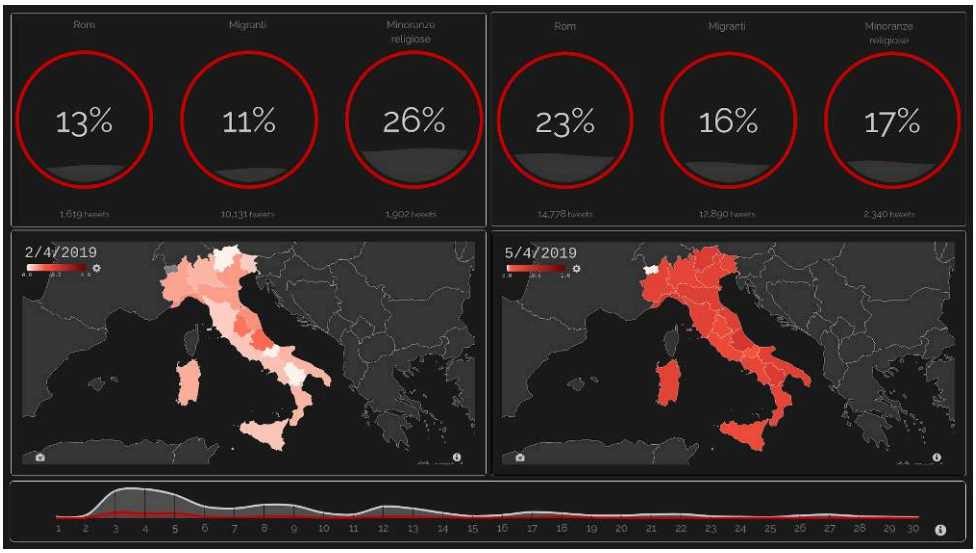

Figure 4
At the bottom, two choropleth maps: on the left, HS is displayed with a linear scale, on the right with a *median scale*. At the top, two examples of liquid fill gauge: on the left, HS percentages for each target on June 3, 2019, on the right, the percentages on June 5.

and Dorling cartogram represent the mean percentage of HS in the tweets created in a region. By default, the color scale is linear between 0 and 1, but the user can switch to a *median scale*, a scale where the median value of the color scale is the median value of HS calculated in all regions in the last 30 days. The Dorling cartogram, unlike the choropleth map, can show a further variable by representing it with the size of the circles. In our visualization platform, the size of each circle is proportional the total number of tweets created in that region.
Liquid Fill Gauge is a circle gauge that represents a percentage value, but in an eye-catching way: we decided to develop also this kind of visualization because it is simple and easy to read.
In Figure 4 the Liquid Fill Gauge shows how 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 suburbs of Rome[17]. 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

The graphical representation of textual information can be challenging, but it is necessary in a platform designed to support the analysis of the hate speech phenomenon also from a computational linguistic point of view. We decided to create two main

17 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

visualizations: (i) visualization of word frequencies; (ii) visualization of co-occurrences among words across tweets.

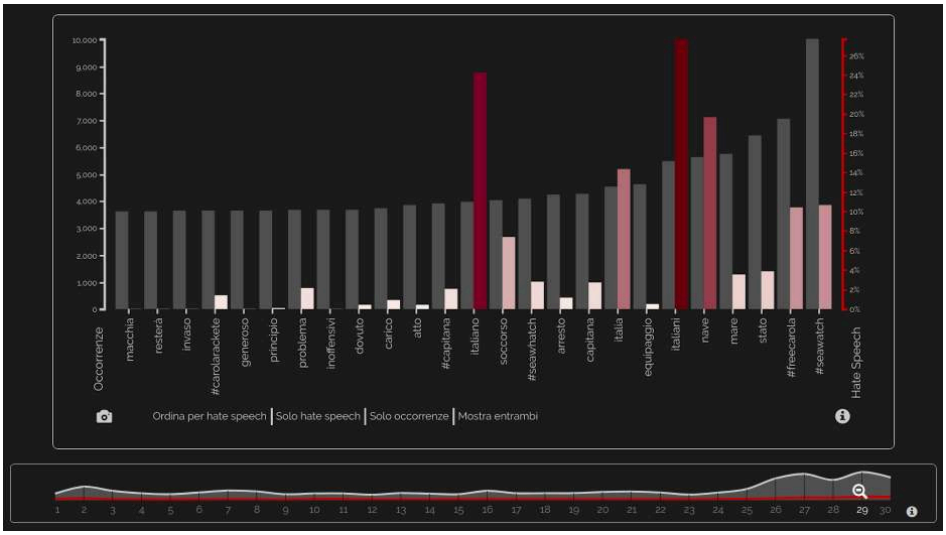

Figure 5
Most frequent words on June 29, 2019. For each word, there are represented the total number of tweets containing that word and the mean percentage of HS in those tweets.

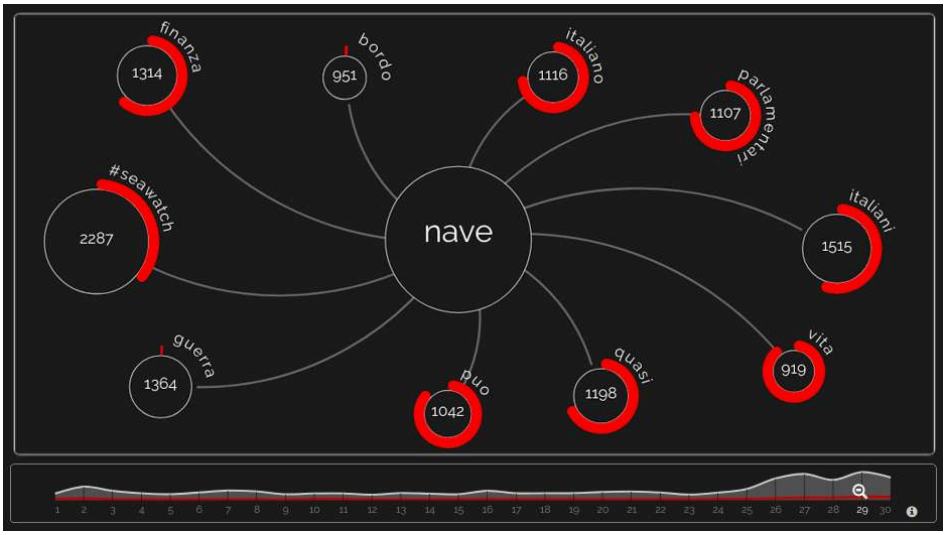

Figure 6
Word co-occurrences network. On June 29, 2019, the word *nave* (ship) co-occurs 2287 times with the hashtag *seawatch*, the NGO ship.

Word frequencies are represented through a bar chart: Figure 5 shows the 25 most occurring words in the selected time period (June 29, 2019). For each word, the chart shows the number of occurrences and the average percentage of HS in tweets containing that word.
Co-occurrences among words are represented as a network. By clicking on a word from

the bar chart also shown in Figure 5, the user can visualize the co-occurrence network of that word (Figure 6).

If we analyze together the *Polarized Weirdness Index* results (see Section 5), and the example of word co-occurences network reported in 6, we notice that often the most characteristic terms in a HS message are not necessarily hate words. In fact, much hatred content is conveyed in an implicit form. Therefore, a deeper study of the semantic shift of recurring words in the public debate is needed. For instance, the widely used metonymy between *barconi* (boats) and immigrants could be perceived as a dehumanization of the target. Another example is the word *italiani* (Italians), which is often used to express a we-others dichotomy. Not only that, the longitudinal analysis presented in Section 3.3 shows the importance of rethinking the criteria we use to collect potential HS. On the one hand, it is possible to reduce the bias related to an arbitrary choice of the words used to filter content (Wiegand, Ruppenhofer, and Kleinbauer 2019). On the other hand, using high-WPW terms in the data gathering stage might lead to a more exhaustive collection of texts, in which also indirect ways of referring to the targets of HS could be analyzed. In very recent work, polarized weirdness is applied to pre-trained word embeddings to encode the semantic shift occurring with domains such as HS, but also for other tasks such as gender prediction (Basile 2020).

4.3 Showing Hater Users

The output of the HATECHECKER component is shown through an interactive user interface.

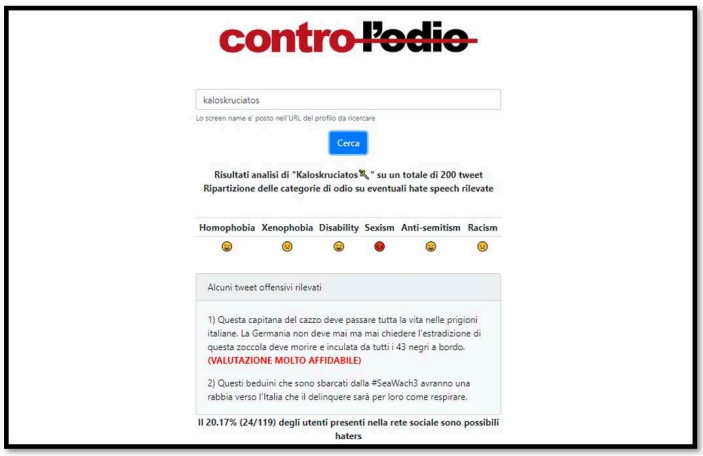

Figure 7
A screenshot of HATECHECKER at work

A screenshot of the working prototype of the platform is reported in Fig. 7. As shown in the figure, a user interacting with the platform can query the system by interactively providing a Twitter user name. In a few seconds, the interface shows a report of the target user containing a set of emojis reporting the behavior of the user for each of the categories we analyzed, a snapshot of their own tweets labeled as hate speech and some information about the percentage of hater profiles that are in the social network of the target user.

It is worth noting that such a web application is very useful for both monitoring tasks (e.g., to verify whether a third-party account is an online hater) as well as for *Quantified Self* scenarios (Swan 2013), that is to say, to improve the self-awareness and the self-consciousness of the user towards the dynamics of her social network. Our intuition is that a user who is aware of not being a hater, can use the system to identify (if any) the haters that are still in her own social network, and maybe decide to unfollow them or, hopefully, to understand the discomfort behind their manifested hostility, and to try to alleviate it if the strength of the relationship allows it. Additionally, it can be used as a tool for self assessment and self observation to check our own level of hate speech in our tweets.

5. Countering Online Hate Speech in High Schools

The interactive hate maps and the 'Words of Hate' visualization settings described here were 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, Rom, and religious minorities, and on the creation of positive narratives to actively counteract hatred online. A team of twenty educators carried out 150 laboratories in seven different Italian regions (Piedmont, Tuscany, Liguria, Emilia Romagna, Lazio, Friuli-Venezia Giulia, and Sardinia). Furthermore, a group of teachers built a community of practice in order to integrate Contro l'Odio tools in their teaching curriculum.

Finally, results of the HS detection engine were shared and discussed with people belonging to minorities. Three focus groups were organized with Muslim, Rom, and Copts, who brought out the most recurring stereotypes referred to them.

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 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.

The "Contro l'Odio" project is also meaningfully impacting on research activities. In particular, a portion of data automatically annotated by the platform has been included in the dataset of the second Hate Speech Detection shared task (HaSpeeDe 2)[18] organized within EVALITA 2020[19](Sanguinetti et al. 2020). Besides the main task on

18 http://www.di.unito.it/~tutreeb/haspeede-evalita20/index.html
19 The 7th evaluation campaign of Natural Language Processing and Speech tools for Italian, http://www.evalita.it/2020

HS classification, two additional pilot tasks have been introduced in this edition, that aim to further explore HS phenomenon under different perspectives. The first one is a classification task aimed to detect the presence of stereotypes referred to the same hate targets included in the Contro l'Odio project. Furthermore, drawing inspiration from previous work on syntactic realization of hateful content (see the POP-HS-IT corpus in Comandini and Patti (2019)), a new sequence labeling task has also been proposed, aiming to identify nominal utterances specifically in hate speech data. For this purpose the dataset has been released with a richer annotation that can in turn be used in the future for other research activities.

Another relevant side effect of the "Contro l'Odio" is the project "Be Positive!"[20]. As a follow up of "Contro l'Odio", "Be Positive!" aims at automatically collecting and identifying online HS in order to increase positive contents (counternarratives) addressed to groups vulnerable to discrimination and promote their active presence on social media. The project involves the improvement of Hate maps developed within "Contro l'Odio" and described above, and the creation of an automatic writing assistant that suggests positive contents against HS, together with the organization of training courses addressed to schools, journalists, communication experts, health care workers, minorities, and activists.

Acknowledgments

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

Ahmad, Khurshid, Lee Gillam, and Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *The Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, MD, USA, January. National Institute of Standards and Technology (NIST).

Basile, Valerio. 2020. Domain adaptation for text classification with weird embeddings. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, March. CEUR-WS.org.

Basile, Valerio, 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, Minneapolis, MN, USA, June. Association of Computational Linguistics.

Basile, Valerio, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Bassignana, Elisa, 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)*, volume 2253 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Bosco, Cristina, 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, Turin, Italy, December. CEUR.

20 "Be Positive!" is funded in 2020-22 under the "Google.org Impact Challenge on Safety" call.

Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July. Association for Computational Linguistics.

Comandini, Gloria and Viviana Patti. 2019. An impossible dialogue! nominal utterances and populist rhetoric in an Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 163–171, Florence, Italy, August. Association for Computational Linguistics.

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

Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August. Association for Computational Linguistics.

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

Del Vigna, Fabio, 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)*, Venice, Italy, January.

EU Commission. 2016. Code of conduct on countering illegal hate speech online.

Fersini, Elisabetta, Debora Nozza, and Paolo Rosso. 2018. 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)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Fersini, Elisabetta, Debora Nozza, and Paolo Rosso. 2020. Ami @ evalita2020: Automatic misogyny identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR.org.

Fersini, Elisabetta, Paolo Rosso, and Maria Anzovino. 2018. 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)*, volume 2150 of *CEUR Workshop Proceedings*, page 214–228, Sevilla, Spain, September. CEUR-WS.org.

Florio, K., V. Basile, M. Lai, and V. Patti. 2019. Leveraging hate speech detection to investigate immigration-related phenomena in italy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–7, Cambridge, UK, September.

Florio, Komal, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).

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

Kullback, S. and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lingiardi, Vittorio, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2020. Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, 39(7):711–721.

Menini, Stefano, 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*, Florence, Italy, August. Association of Computational Linguistics.

Musto, Cataldo, 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, Halifax, Nova Scotia, Canada, July. ACM.

Pamungkas, Endang Wahyu, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6237–6246, Marseille, France, May. European Language Resources Association.

Paschalides, Demetris, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. 2020. Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Transactions on Internet Technology*, 20(2), March.

Pedersen, Ted. 2010. Information content measures of semantic similarity perform better without sense-tagged text. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 329–332, Los Angeles, CA, USA, June. Association for Computational Linguistics.

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and Benchmark Corpora for Hate Speech Detection: a Systematic Review. *Language Resources and Evaluation*.

Poletto, Fabio, 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. CEUR.

Polignano, Marco, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019a. A Comparison of Word-Embeddings in Emotion Detection from Text using BiLSTM, CNN and Self-Attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68, Larnaca, Cyprus, June.

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

Sanguinetti, Manuela, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2@evalita2020: Overview of the evalita 2020 hate speech detection task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR.org.

Sanguinetti, Manuela, 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. European Language Resources Association (ELRA).

Schmidt, Anna 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*, Valencia, Spain, April. Association for Computational Linguistics.

Swan, Melanie. 2013. The quantified self: Fundamental disruption in big data science and biological discovery. *Big data*, 1(2):85–99.

Vilella, Salvatore, Mirko Lai, Daniela Paolotti, and Giancarlo Ruffo. 2020. Immigration as a divisive topic: Clusters and content diffusion in the italian twitter debate. *Future Internet*, 12(10):173.

Wiegand, Michael, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, MN, USA, June.