HATECHECKER: a Tool to Automatically Detect *Hater* Users in Online Social Networks

Cataldo Musto
University of Bari
Dip. di Informatica
cataldo.musto@uniba.it

Angelo Pio Sansonetti
University of Bari
Dip. di Informatica (*Bachelor Student*)
a.sansonetti6@studenti.uniba.it

Marco Polignano
University of Bari
Dip. di Informatica
marco.polignano@uniba.it

Giovanni Semeraro
University of Bari
Dip. di Informatica
giovanni.semeraro@uniba.it

Marco Stranisci
Associazione ACMOS
Torino
marco.stranisci@acmos.net

Abstract

In this paper we present HATECHECKER, a tool for the automatic detection of *hater* users in online social networks which has been developed within the activities of "Contro L'Odio" research project.

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 speeches*. *(iii)* a lexicon is used to classify hate speeches against a set of specific categories that can describe the target user (*e.g.,* racist, homophobic, antisemitic, etc.).

Finally, the output of the tool, that is to say, a set of labels describing (if any) the intolerant traits of the target user, are shown through an interactive user interface and exposed through a REST web service for the integration in third-party applications.

In the experimental evaluation we crawled and annotated a set of 200 Twitter profiles and we investigated to what extent our tool is able to correctly identify *hater* users. The results confirmed the validity of our methodology and paved the way for several future research directions.

1 Background and Motivations

According to a recent study[1], 58% of the Italian population regularly uses online social networks as Twitter, Facebook, Instagram and LinkedIn.

Such a huge diffusion of these platforms is providing the users with many new opportunities and services, just think that almost everyone now uses social media to get information, discuss, express opinions and stay in touch with friends. Unfortunately, due to the lack of control and the absence of a clear management of the concept of *identity* of the users, social networks have become the *perfect place* to spread hate against minorities and people having different cultures, values and opinions.

As pointed out by several works (Mathew et al., 2018), the diffusion of *hate speeches* in online social media is continuously growing and the countermeasures adopted by the single platforms are neither effective nor timely, even if a big effort is done to make the process of removing hate speeches faster and more precise[2]. Accordingly, the research line related to the development of tools and methods for the *automatic detection of hate speeches* gained more and more attention. Techniques for detecting hate speeches are obviously based on NLP techniques, and range from simple lexicon-based approaches (Gitari et al., 2015) to more sophisticated techniques that exploit word embeddings (Djuric et al., 2015) and deep learning methods (Badjatiya et al., 2017).

Similar research attempts were also proposed for the Italian language. One of the most popular initiative is the Italian HateMap project (Musto

[1] https://wearesocial.com/it/blog/2018/01/global-digital-report-2018

[2] https://www.cnbc.com/2019/02/04/facebook-google-and-twitter-are-getting-faster-at-removing-hate-speech-online-eu-finds–.html

et al., 2016), a research project that exploits semantic analysis and opinion mining to identify the most-at-risk areas of the Italian country, that is to say, the areas where the users more frequently publish hate speeches. The interest of the research community for the topic was confirmed by the recent work by Bosco et al. (Bosco et al., 2017), who studied hate speech against immigrants, and by Anzovino et al. (Anzovino et al., 2018) who detected misogyny on Twitter. Moreover, as shown by the organization of a specific task in the EVALITA evaluation campaign, an important effort is now devoted to the automatic detection of misogyny (Fersini et al., 2018) and hate speeches in general (Bosco et al., 2018; Basile et al., 2019).

In order to continue the investigation in this research line ACMOS[3], a no-profit association based in Torino, recently launched "Contro l'Odio[4]", a joint research project with the University of Bari, University of Torino and several local associations. The project aims to develop tools and methodologies to monitor (and *hopefully* tackle) online hate speeches and intolerant behaviors.

One of the outcomes of the research is HATECHECKER, a tool that aims to automatically identify *hater* users on Twitter by exploiting sentiment analysis and natural language processing techniques. The distinguishing aspect of the tool with respect to the work we have previously introduced is the *focus* of the tool itself. Indeed, differently from most of the literature, that focused on the analysis of single Tweets, HATECHECKER aims to analyze the users *as a whole*, and to identify *hater users* rather than *hate speeches*. Clearly, both the tasks are in close correlation, since techniques to detect hate speeches can be used to detect hater users as well.

However, through this work we want to move the focus on the latter since, up to our knowledge, this a poorly investigated research direction. Just think that no datasets of *hater users* is currently publicly available.

To sum up, the contributions of the work can be summarized as follows:

- We present a workflow that allows to detect *hater users* in online social networks;
- We evaluate several configurations (on varying of lexicons and sentiment analysis algorithms) of the pipeline and we identified the most effective one to tackle our specific task;
- We share the first publicly available dataset for automatic detection of *hater users* on Twitter.

In the following, we will first describe the methodology we designed to implement our system, then we will discuss the effectiveness of the approach by analyzing the results we obtained on a (publicly available) dataset of 200 Twitter users.

2 Methodology

The workflow carried out by the HATECHECKER tool is reported in Figure 1.

Generally speaking, the pipeline consists of four different modules, that is to say, a SOCIAL DATA EXTRACTOR, a SENTIMENT ANALYZER, a PROFILE CLASSIFIER anda SOCIAL NETWORK PROCESSOR. All these components use a `NoSQL` database to store the information they hold and expose the output returned by the tool through a `REST` interface as well as through a Web Application. In the following, a description of the single modules that compose the workflow is provided.

2.1 Social Data Extractor

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 absence of textual content, it is not possible provide such a classification. To this end, the first and mandatory step carried out by the tool is the extraction of the Tweets posted by the user we want to analyze. In this case, we used the official Twitter APIs to gather the available Tweets and to forward it to the next modules of the workflow.

Given that the *real-time execution* of the workflow is one of the constraints of the project, we limited the extraction to the 200 most recent Tweets posted by the user. This is a reasonable choice, since we aim to detect users who *recently* showed an intolerant behavior, rather than users who posted hate speeches one or two years ago.

2.2 Sentiment Analyzer

Once the Tweets have been collected, it is necessary to provide the tool with the ability to go

[3]http://www.acmos.net
[4]http://www.controlodio.it

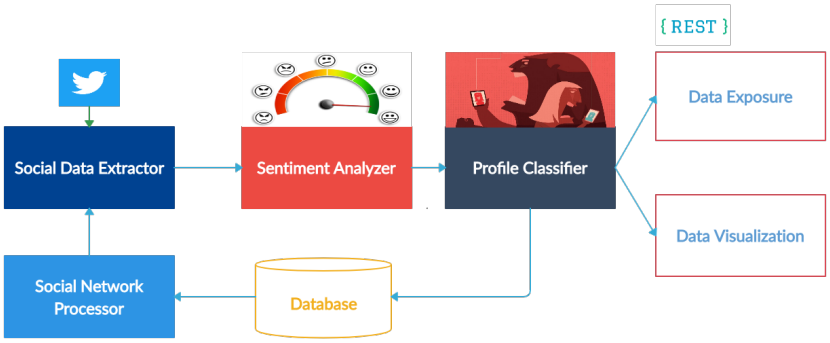

Figure 1: The workflow carried out by the HATECHECKER tool

through the content posted by the target and to automatically identify the *hate speeches*.

To this end, the SENTIMENT ANALYZER modules exploits Sentiment Analysis techniques (Pang et al., 2008) to basically classify each Tweet as *positive* or *negative* (that it to say, conveying hate speeches or not). To get this output we integrated and compared two different implementations of sentiment analysis algorithms:

- **SentiPolC:** (Basile and Novielli, 2014) a sentiment analysis algorithm that resulted as the best-performing one in EVALITA 2014 in carrying out the task of associating the correct *sentiment* to Tweets;
- **HanSEL:** an algorithm based on a deep neural network C-BiLSTN (Zhou et al., 2015) with an input layer of word embeddings. This strategy is based on the work proposed by Polignano et al. (Polignano and Basile, 2018) and it has been improved within the activities of the 'Contro l'Odio' research project. In particular, the whole net has been trained for 20 epochs with early stopping criteria, Adam loss function, and binary cross-entropy as optimization function.

A complete overview of the algorithms is out of the scope of this paper and we suggest to go through the references for a thorough discussion. For the sake of simplicity, we can state that the output of both the algorithms is a *binary* classification of each Tweet posted by the target user as *negative* (that is to say, conveying hate speeches) or *positive*. Such an output is then passed to the PROFILE CLASSIFIER module whose goal is to assign a more precise label to the user, on the ground of the nature of the hate speeches she posted (if any).

2.3 Profile Classifier

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, the answers to two fundamental questions are still lacking:

- How can we label the user *as hater or non-hater* on the ground of the Tweets she posted?
- How can we return a more fine-grained classification of the user (*e.g.,* racist, homofobe, etc.) on the ground of the Tweets she posted?

Both these issues are tackled by the PROFILE CLASSIFIER module. As for the first question, a very simple strategy based on *thresholding* is implemented. In particular, we defined a parameter ϵ, and whether the user posted a number of Tweets labeled as *hate speeches* higher than ϵ, the user herself is labeled as an *hater*. Of course, several values for the parameter ϵ can be taken into account to run the tool.

As for the second question, we used a *lexicon-based* approach to provide a fine-grained classification of users' profiles. The intuition behind our methodology is that for each category a specific lexicon can be defined, and whether 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.

To define the lexicon for each category, we relied on the research results of the Italian Hate Map

(Lingiardi et al., 2019). In particular, we exploited the categories as well as the lexicon used in the Italian Hate Map Project, which consists of 6 differents 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, a multilingual lexicon of hate words (Bassignana et al., 2018). 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.

Obviously, in the experimental session the effectiveness of the tool on varying of different lexicons and on different configuration of the workflow will be evaluated.

2.4 Social Network Processor

At the end of the previous step, the target user is labeled with a set of categories describing the *facets* of her 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, of course) 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*.

Unfortunately, due to space reasons, the discussion of this part of the workflow is out of the scope of this paper and is left for future discussions.

2.5 Data Exposure and Data Visualization

Finally, the output of the platform is made available to third-party services and to the user itself. In the first case, a REST web service makes available the output of the tool (that it so say, the hate categories and the number of haters in her own social network), while in the latter the same data are shown through an interactive user interface.

A screenshot of the working prototype of the platform is reported in Figure 2. 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 her own Tweets labeled as hate speeches and some information about the percentage of hater profiles that are in the social network of the target user.

It is worth to note 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 an 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.

3 Experimental Evaluation

The goal of the experimental session was to evaluate the effectiveness of the tool on varying of different configurations of the pipeline.

To this end, due to the lack of a dataset of *hater profiles*, we manually crawled and annotated a set of 200 Twitter users, which we made available[5] for the sake of reproducibility and to foster the research in the area.

In particular, we compared four different strategies to run our tool, on varying on two different parameters, such as the lexicon and the sentiment analysis algorithm. In particular, we exploited the following combinations of parameters:

- **Sentiment Analysis:** SentipolC and HanSEL, as previously explained
- **Lexicons:** HateMap lexicon and complete lexicon (HateMap+Hurtlex)

As for the parameters, the threshold ϵ was set equal to 3 and both the sentiment analysis algorithms were run with the standard parameters introduced in the original papers. To evaluate the effectiveness of the approaches, we calculate the number of correctly classified user profiles over the total of hater users in the dataset.

[5] https://tinyurl.com/uniba-haters-dataset

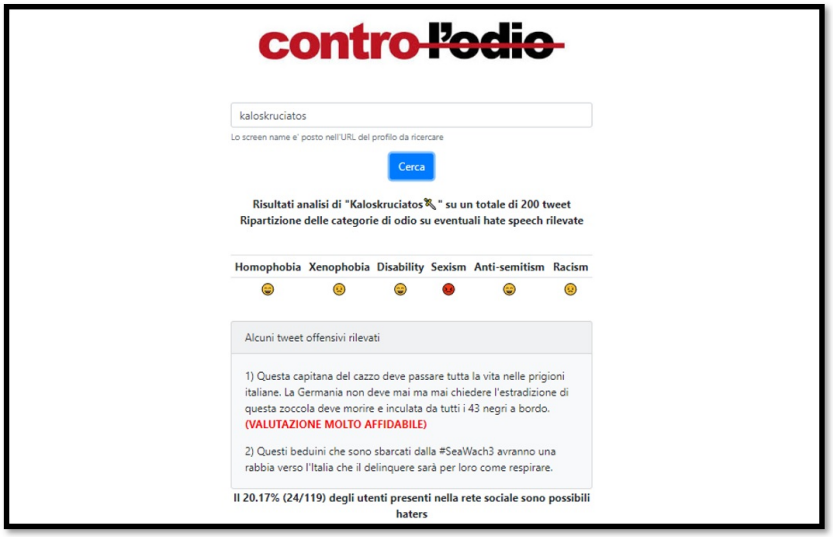

Figure 2: A screenshot of HATECHECKER at work

		Facets					
Lexicon	**Algorithm**	*Racism*	*Anti-semitism*	*Disability*	*Sexism*	*Homophobia*	*Xenophobia*
HateMap	SentiPolC	71.5	92.0	82.0	77.5	**84.0**	75.5
HateMap	HanSEL	73.0	95.5	**88.5**	**79.0**	**84.0**	**79.0**
Complete	SentiPolC	**78.0**	95.0	86.5	78.0	**84.0**	78.0
Complete	HanSEL	75.0	**97.0**	**88.5**	78.0	**84.0**	**79.0**

Table 1: Results of the Experiment. The best-performing configuration for each facet is reported in **bold**.

The results of the experiments are reported in Table 1. In general, we can state that our approach to automatically detect hater users in online social network provided us with encouraging results, since more a percentage between 78% and 97% of the online haters were correctly detected by the algorithm, regardless of the specific category.

It is worth to note that the *worse* results (both of them are beyond 70%, through) were obtained for *racism* and *xenophobia*, that is to say, two facets characterized by a lexicon that quickly evolves and often adopts terms that are *not conventional* and not necessarily conveying *hate* (*e.g.*,, expressions as *'Aiutiamoli a casa loro'* or terms as *'clandestini'*). However, even for these categories the results we obtained were encouraging.

Conversely, results were particularly outstanding for facets such as *anti-semitism* and *homophobia*, that have a quite fixed *lexicon* of terms that can be used to hurt or offend such minorities.

As for the different configurations, we noted that HANSEL tended to obtain better results than SENTIPOLC. This is a *quite* expected outcome, since it exploits more novel and effective methods as those based on word embeddings and deep learning techniques. Moreover, we can state that the results can be further improved since no particular tuning of the parameters was carried out in this work.

As for the lexicons, the extension of the original Italian Hate Map lexicons with new terms led to an improvement of the results for all the facets (except for *homophobia*) for at least one of the comparisons. Such improvement are often tiny, but this is an expected outcome since just a few terms coming from Hurtlex were added. However, even these preliminary results provided us with encouraging findings, since they showed that the integration and the extension of sensible terms with the information coming from recently developed lexical resources can lead to a further improvement of the accuracy of the system.

4 Conclusions and Future Work

In this work we have presented HATECHECKER, a tool that exploits sentiment analysis and natural language processing techniques to automatically detect *hater users* in online social networks.

Given a target user, the workflow we implemented in our system uses sentiment analysis techniques to identify hate speeches posted by the user and exploits a lexicon that extends that of the Italian Hate Map project to assign to the person one or more labels that describe the nature of the hate speeches she posted.

As future work, we plan to arrange a user study, specifically designed for *young people*, to evaluate the effectiveness of the system as a Quantified Self tool (Musto et al., 2018), that is to say, to improve the awareness of the users towards the behavior of other people in their social network.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Pierpaolo Basile and Nicole Novielli. 2014. Uniba at evalita 2014-sentipolc task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. *Proceedings of EVALITA*, pages 58–63.

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

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Cristina Bosco, Viviana Patti, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Francesco Ruffo, Rossano Schifanella, and Marco Stranisci. 2017. Tools and resources for detecting hate and prejudice against immigrants in social media. In *Symposium III. Social Interactions in Complex Intelligent Systems (SICIS) at AISB 2017*, pages 79–84. AISB.

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

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@ CLiC-it*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

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

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2018. Spread of hate speech in online social media. *arXiv preprint arXiv:1812.01693*.

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

Cataldo Musto, Giovanni Semeraro, Cosimo Lovascio, Marco de Gemmis, and Pasquale Lops. 2018. A framework for holistic user modeling merging heterogeneous digital footprints. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 97–101. ACM.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Marco Polignano and Pierpaolo Basile. 2018. Hansel: Italian hate speech detection through ensemble learning and deep neural networks. In *EVALITA@ CLiC-it*.

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

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*.