Overview of the EVALITA 2018 Hate Speech Detection Task

Cristina Bosco
University of Torino
Italy
bosco@di.unito.it

Felice Dell'Orletta
ILC-CNR, Pisa
Italy
felice.dellorletta@ilc.cnr.it

Fabio Poletto
Acmos, Torino
Italy
fabio.poletto@edu.unito.it

Manuela Sanguinetti
University of Torino
Italy
msanguin@di.unito.it

Maurizio Tesconi
IIT-CNR, Pisa
Italy
maurizio.tesconi@iit.cnr.it

Abstract

English. The Hate Speech Detection (HaSpeeDe) task is a shared task on Italian social media (Facebook and Twitter) for the detection of hateful content, and it has been proposed for the first time at EVALITA 2018. Providing two datasets from two different online social platforms differently featured from the linguistic and communicative point of view, we organized the task in three tasks where systems must be trained and tested on the same resource or using one in training and the other in testing: HaSpeeDe-FB, HaSpeeDe-TW and Cross-HaSpeeDe (further subdivided into Cross-HaSpeeDe_FB and Cross-HaSpeeDe_TW sub-tasks). Overall, 9 teams participated in the task, and the best system achieved a macro F1-score of 0.8288 for HaSpeeDe-FB, 0.7993 for HaSpeeDe-TW, 0.6541 for Cross-HaSpeeDe_FB and 0.6985 for Cross-HaSpeeDe_TW. In this report, we describe the datasets released and the evaluation measures, and we discuss results.

Italiano. *HaSpeeDe è la prima campagna di valutazione di sistemi per l'identificazione automatica di discorsi di incitamento all'odio su social media (Facebook e Twitter) in lingua italiana, proposta nell'ambito di EVALITA 2018. Fornendo ai partecipanti due insiemi di dati estratti da due piattaforme differenti dal punto di vista linguistico e della comunicazione, abbiamo articolato HaSpeeDe in tre compiti in cui i sistemi sono addestrati e testati sulla stessa tipologia di dati oppure addrestrati su una tipologia e testati sull'altra: HaSpeeDe-FB, HaSpeeDe-TW e Cross-HaSpeeDe (a sua volta suddiviso in Cross-HaSpeeDe_FB e Cross-HaSpeeDe_TW). Nel complesso, 9 gruppi hanno partecipato alla campagna, e il miglior sistema ha ottenuto un punteggio di macro F1 pari a 0,8288 in HaSpeeDe-FB, 0,7993 in HaSpeeDe-TW, 0,6541 in Cross-HaSpeeDe_FB e 0.6985 in Cross-HaSpeeDe_TW. L'articolo descrive i dataset rilasciati e le modalità di valutazione, e discute i risultati ottenuti.*

1 Introduction and Motivations

Online hateful content, or Hate Speech (HS), is characterized by some key aspects (such as virality, or presumed anonymity) which distinguish it from offline communication and make it potentially more dangerous and hurtful. Therefore, its identification becomes a crucial mission in many fields.

The task that we have proposed for this edition of EVALITA namely consists in automatically annotating messages from two popular microblogging platforms, Twitter and Facebook, with a boolean value indicating the presence (or not) of HS.

HS can be defined as any expression "*that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth*" (Erjavec and Kovačič, 2012).

Although definitions and approaches to HS vary a lot and depend on the juridical tradition of the country, many agree that what is identified as

such can not fall under the protection granted by the right to freedom of expression, and must be prohibited. Also for transposing in practical initiatives the Code of Conduct of the European Union[1], online platforms like Twitter, Facebook or YouTube discourage hateful content, but its removal mainly relies on users and trusted flaggers reports, and lacks a systematic control.

Although HS analysis and identification requires a multidisciplinary approach that includes knowledge from different fields (psychology, law, social sciences, among others), NLP plays a fundamental role in this respect. Therefore, the development of high-accuracy automatic tools able to identify HS assumes the utmost relevance not only for NLP – and Italian NLP in particular – but also for all the practical applications a similar task lends itself to. Furthermore, as also suggested in Schmidt and Wiegand (2017), the community would considerably benefit from a benchmark dataset for HS detection underlying a commonly accepted definition of the task.

As regards the state of the art, a large number of contributions have been proposed on this topic, that adopt from lexicon-based (Gitari et al., 2015) to various machine learning approaches, and with different learning techniques, ranging from naïve Bayes classifiers (Kwok and Wang, 2013), Logistic Regression and Support Vector Machines (Burnap and Williams, 2015; Davidson et al., 2017), to the more recent Recurrent and Convolutional Neural Networks (Mehdad and Tetreault, 2016; Gambäck and Sikdar, 2017). However, there exist no comparative studies which would allow making judgement on the most effective learning method (Schmidt and Wiegand, 2017).
Furthermore, a large number of academic events and shared tasks took place in the recent past, thus reflecting the interest in HS and HS-related topics by the NLP community; to name a few, the first and second edition of the Workshop on Abusive Language[2] (Waseem et al., 2017), the First Workshop on Trolling, Aggression and Cyberbullying (Kumar et al., 2018), that also included a shared task on aggression identification, the tracks on Automatic Misogyny Identification (AMI) (Fersini et al., 2018b) and on authorship and aggressiveness analysis (MEX-A3T) (Carmona et al., 2018) proposed at the 2018 edition of IberEval, the GermEval Shared Task on the Identification of Offensive Language (Wiegand et al., 2018), the Automatic Misogyny Identification task at EVALITA 2018 (Fersini et al., 2018a), and finally the SemEval shared task on hate speech detection against immigrants and women (HatEval), that is still ongoing at the time of writing[3].

On the other hand, such contributions and events are mainly based on other languages (English, for most part), while very few of them deal with Italian (Del Vigna et al., 2017; Musto et al., 2016; Pelosi et al., 2017). Precisely for this reason, the Hate Speech Detection (HaSpeeDe)[4] task has been conceived and proposed within the EVALITA context (Caselli et al., 2018); its purpose is namely to encourage and promote the participation of several research groups, both from academia and industry, making a shared dataset available, in order to allow an advancement in the state of the art in this field for Italian as well.

2 Task Organization

Considering the linguistic, as well as metalinguistic, features that distinguish Twitter and Facebook posts, namely due to the differences in use between the two platforms and the character limitations posed for their messages (especially on Twitter), the task has been further organized into three sub-tasks, based on the dataset used (see Section 3):

- **Task 1: HaSpeeDe-FB**, where only the Facebook dataset could be used to classify the Facebook test set
- **Task 2: HaSpeeDe-TW**, where only the Twitter dataset could be used to classify the Twitter test set
- **Task 3: Cross-HaSpeeDe**, which has been further subdivided into two sub-tasks:
  - **Task 3.1: Cross-HaSpeeDe_FB**, where only the Facebook dataset could be used to classify the Twitter test set
  - **Task 3.2: Cross-HaSpeeDe_TW**, where, conversely, only the Twitter

[1] On May 31, 2016, the EU Commission presented with Facebook, Microsoft, Twitter and YouTube a "Code of conduct on countering illegal hate speech online".

[2] https://sites.google.com/view/alw2018/

[3] https://competitions.codalab.org/competitions/19935

[4] http://www.di.unito.it/~tutreeb/haspeede-evalita18/

dataset could be used to classify the Facebook test set

Cross-HaSpeeDe, in particular, has been proposed as an out-of-domain task that specifically aimed on one hand at highlighting the challenging aspects of using social media data for classification purposes, and on the other at enhancing the systems' ability to generalize their predictions with different datasets.

3 Datasets and Format

The datasets proposed for this task are the result of a joint effort of two research groups on harmonizing the annotation previously applied to two different datasets, in order to allow their exploitation in the task.

The first dataset is a collection of Facebook posts developed by the group from Pisa and created in 2016 (Del Vigna et al., 2017), while the other one is a Twitter corpus developed in 2017-2018 by the Turin group (Sanguinetti et al., 2018). Section 3.1 and 3.2 briefly introduce the original datasets, while Section 3.3 describes the unified annotation scheme adopted in both corpora for the purposes of this task.

3.1 Facebook Dataset

This is a corpus of comments retrieved from the Facebook public pages of Italian newspapers, politicians, artists, and groups. Those pages were selected because typically they host discussions spanning across a variety of topics.

The comments collected were related to a series of web pages and groups, chosen as being suspected to possibly contain hateful content: *salviniofficial*, *matteorenziufficiale*, *lazanzarar24*, *jenusdinazareth*, *sinistracazzateliberta2*, *ilfattoquotidiano*, *emosocazzi*, *noiconsalviniufficiale*.

Overall, 17,567 Facebook comments were collected from 99 posts crawled from the selected pages. Five bachelor students were asked to annotate comments, in particular 3,685 received at least 3 annotations. The annotators were asked to assign one class to each post, where classes span over the following levels of hate: *No hate*, *Weak hate*, *Strong hate*.

Hateful messages were then divided into distinct categories: *Religion*, *Physical and/or mental handicap*, *Socio-economical status*, *Politics*, *Race*, *Sex and Gender issues*, and *Other*.

3.2 Twitter Dataset

The Twitter dataset released for the competition is a subset of a larger hate speech corpus developed at the Turin University. The corpus forms indeed part of the Hate Speech Monitoring program[5], coordinated by the Computer Science Department with the aim at detecting, analyzing and countering HS with an inter-disciplinary approach (Bosco et al., 2017). Its preliminary stage of development has been described in Poletto et al. (2017), while the fully developed corpus is described in Sanguinetti et al. (2018).

The collection includes Twitter posts gathered with a classical keyword-based approach, more specifically by filtering the corpus using neutral keywords related to three social groups deemed as potential HS targets in the Italian context: immigrants, Muslims and Roma.

After a first annotation step that resulted in a collection of around 1,800 tweets, the corpus has been further expanded by adding new annotated data. The newly introduced tweets were annotated partly by experts and partly by CrowdFlower (now Figure Eight) contributors. The final version of the corpus consists of 6,928 tweets.

The main feature of this corpus is its annotation scheme, specifically designed to properly encode the multiplicity of factors that can contribute to the definition of a hate speech notion, and to offer a broader tagset capable of better representing all those factors which may increase, or rather mitigate, the impact of the message. This resulted in a scheme that includes, besides HS tags (*no-yes*), also its intensity degree (from *1* through *4* if HS is present, and *0* otherwise), the presence of aggressiveness (*no-weak-strong*) and offensiveness (*no-weak-strong*), as well as irony and stereotype (*no-yes*).

In addition, given that irony has been included as annotation category in the scheme, part of this hate speech corpus (i.e. the tweets annotated as ironic) has also been used in another task proposed in this edition of EVALITA, namely the one on irony detection in Italian tweets (IronITA)[6](Cignarella et al., 2018). More precisely, the overlapping tweets in the IronITA datasets are 781 in the training set and just 96 in the test set.

[5] http://hatespeech.di.unito.it/

[6] http://www.di.unito.it/~tutreeb/ironita-evalita18/

3.3 Format and Data in HaSpeeDe

The annotation format provided for the task is the same for both datasets described above, and it consists of a simplified version of the schemes adopted in the two corpora introduced in Section 3.1 and 3.2.

The data have been encoded in UTF-8 plain-text files with three tab-separated columns, each one representing the following information:

1. the ID of the Facebook comment or tweet[7],
2. the text,
3. the class: `1` if the text **contains** HS, and `0` otherwise (see Table 1 and 2 for a few examples).

id	text	hs
8	*Io voterò NO NO E NO*	0
36	*Matteo serve un colpo di stato. Qua tra poco dovremo andare in giro tutti armati come in America.*	1

Table 1: Annotation examples from the Facebook dataset.

id	text	hs
1,783	*Corriere: Mafia Capitale, 4 patteggiamenti Gli appalti truccati dei campi rom*	0
3,290	*altro che profughi? sono zavorre e tutti uomini*	1

Table 2: Annotation examples from the Twitter dataset.

Both Facebook and Twitter datasets consist of a total amount of 4,000 comments/tweets retrieved from the main corpora introduced in Section 3.1 and 3.2. The data were randomly split into development and test set, of 3,000 and 1,000 messages respectively.
The distribution in both datasets of the labels expressing the presence or not of HS is summarized in Table 3 and 4.

	0	1
Train	1,618	1,382
Test	323	677
total	1,941	2,059

Table 3: Label distribution in the Facebook dataset.

	0	1
Train	2,028	972
Test	676	324
total	2,704	1,296

Table 4: Label distribution in the Twitter dataset.

4 Evaluation

Participants were allowed to submit up to 2 runs for each task, and a separate official ranking has been provided.
The evaluation has been performed according to the standard metrics known in literature, i.e Precision, Recall and F1-score. However, given the imbalanced distribution of hateful *vs* not hateful messages, and in order to get more useful insights on the system's performance on a given class, the scores have been computed for each class separately; finally the F1-score has been macro-averaged, so as to get the overall results.

For all tasks, the baseline score has been computed as the performance of a classifier based on the most frequent class.

5 Overview of the Task: Participation and Results

5.1 Task Participants and Submissions

A total amount of 9 teams[8] participated in at least one of the three HaSpeeDe main tasks. Table 5 provides an overview of the teams and their affiliation.

Except for one case, where one run was sent for HaSpeeDe-TW only, all teams submitted at least one run for *all* the tasks.

5.2 Systems

As participants were allowed to submit up to 2 runs for each task, several training options were adopted in order to properly classify the texts.
Furthermore, unlike other tasks, we have chosen to not establish any distinction between *constrained* and *unconstrained* runs, and to allow participants to use all the additional resources that

[7] In order to meet the GDPR requirements, texts have been pseudonymized replacing all original IDs in both datasets with newly-generated ones.

[8] In fact, 11 teams submitted their results, but one team withdrew its submissions, and another one's submissions have been removed from the official rankings by the task organizers.

Team	Affiliation
GRCP	Univ. Politècnica de València + CERPAMID, Cuba
InriaFBK	Univ. Côte d'Azur, CNRS, Inria + FBK, Trento
ItaliaNLP	ILC-CNR, Pisa + Univ. of Pisa
Perugia	Univ. for Foreigners of Perugia + Univ. of Perugia + Univ. of Florence
RuG	University of Groningen + Univ. degli Studi di Salerno
sbMMP	Zurich Univ. of Applied Sciences
StopPropagHate	INESC TEC + Univ. of Porto + Eurecat, Centre Tecn. de Catalunya
HanSEL	University of Bari Aldo Moro
VulpeculaTeam	University of Perugia

Table 5: Participants overview.

they deemed useful for the task (other annotated resources, lexicons, pre-trained word embeddings, etc.), on the sole condition that these were explicitly mentioned in their final report.

Table 6 summarizes the external resources (if any) used by participants to enhance their systems' performance, while the remainder of this section offers a brief overview of the teams' systems and core methods adopted to participate in the task .

GRCP (De la Peña Sarracén et al., 2018) The authors proposed a bidirectional Long Short-Term Memory Recurrent Neural Network with an Attention-based mechanism that allows to estimate the importance of each word; this context vector is then used with another LSTM model to estimate whether a text is hateful or not.

HanSEL (Polignano and Basile, 2018) The system proposed is based on an ensemble of three classification strategies, mediated by a majority vote algorithm: Support Vector Machine with RBF kernel, Random Forest and Deep Multilayer Perceptron. The input social media text is represented as a concatenation of word2vec sentence vectors and a TF-IDF bag of words.

InriaFBK (Corazza et al., 2018) The authors implemented three different classifier models, based on recurrent neural networks, n-gram based models and linear SVC.

ItaliaNLP (Cimino et al., 2018) Participants tested three different classification models: one based on linear SVM, another one based on a 1-layer BiLSTM and a newly-introduced one based on a 2-layer BiLSTM which exploits multi-task learning with additional data from the 2016 SENTIPOLC task (Barbieri et al., 2016).

Perugia (Santucci et al., 2018) The participants' system uses a document classifier based on a SVM algorithm. The features used by the system are a combination of features extracted using mathematical operations on FastText word embeddings and other 20 features extracted from the raw text.

RuG (Bai et al., 2018) The authors proposed two different classifiers: a SVM based on linear kernel algorithm and an ensemble system composed of a SVM classifier and a Convolutional Neural Network combined by a logistic regression meta-classifier. The features of each classifier is algorithm dependent and exploits word embeddings, raw text features and lexical resources features.

sbMMMP The authors tested two different systems, in a similar fashion to what described in von Grüningen et al. (2018). The first one is based on an ensemble of Convolutional Neural Networks (CNN), whose outputs are then used as features by a meta-classifier for the final prediction. The second system uses a combination of a CNN and a Gated Recurrent Unit (GRU) together with a transfer-learning approach based on pre-training with a large, automatically-translated dataset.

StopPropagHate (Fortuna et al., 2018) The authors use a classifier based on Recurrent Neural Networks with a binary cross-entropy as loss function. In their system, each input word is represented by a 10000-dimensional vector which is a one-hot encoding vector.

VulpeculaTeam (Bianchini et al., 2018) According to the description provided by participants, a neural network with three hidden layers was used, with word embeddings trained on a set of previously extracted Facebook comments.

5.3 Results and Discussion

In Table 7, 8, 9 and 10, we report the final results of HaSpeeDe, separated according to the respective sub-task and ranked by the macro F1-score (as described in Section 4)[9].

[9]Due to space constraints, the complete evaluation for all classes has been made available here: `https://goo.gl/xPyPRW`

Team	External Resources
GRCP	pre-trained word embeddings
InriaFBK	emotion lexicon
ItaliaNLP Lab	polarity and subjectivity lexicons + 2 word-embedding lexicons
Perugia	Twitter corpus + hate speech lexicon + polarity lexicon
RuG	pre-trained word embeddings + bad/offensive word lists
sbMMP	pre-trained word embeddings
StopPropagHate	–
HanSEL	pre-trained word embeddings
VulpeculaTeam	polarity lexicon + lists of bad words + pre-trained word embeddings

Table 6: Overview of the additional resources used by participants, besides the datasets provided by the task organizers.

In case of multiple runs, the suffixes "_1" and "_2" have been appended to each team name, in order to distinguish the run number of the submitted file. Furthermore, some of the runs in the tables have been marked with *: this means that they were re-submitted because of file incompatibility with the evaluation script or other minor issues that did not affect the evaluation process.

Team	Macro F1-score
baseline	0.2441
ItaliaNLP_2	0.8288
ItaliaNLP_1	0.8106
InriaFBK_1	0.8002
InriaFBK_2	0.7863
Perugia_2	0.7841
RuG_1	0.7751
HanSEL	0.7738
VulpeculaTeam*	0.7554
RuG_2	0.7428
GRCP_2	0.7147
GRCP_1	0.7144
StopPropagHate_2*	0.6532
StopPropagHate_1*	0.6419
Perugia_1	0.2424

Table 7: Results of the HaSpeeDe-FB task.

In absolute terms, i.e. based on the score of the first-ranked team, the best results have been achieved in the HaSpeeDe-FB task, with a macro F1 of 0.8288, followed by HaSpeeDe-TW (0.7993), Cross-HaSpeeDe_TW (0.6985) and Cross-HaSpeeDe_FB (0.6541).

The robustness of an approach benefiting from a polarity and subjectivity lexicon is confirmed by the fact that the best ranking team in both

Team	Macro F1-score
baseline	0.4033
ItaliaNLP_2	0.7993
ItaliaNLP_1	0.7982
RuG_1	0.7934
InriaFBK_2	0.7837
sbMMMP	0.7809
InriaFBK_1	0.78
VulpeculaTeam*	0.7783
Perugia_2	0.7744
RuG_2	0.753
StopPropagHate_2*	0.7426
StopPropagHate_1*	0.7203
GRCP_1	0.6638
GRCP_2	0.6567
HanSEL	0.6491
Perugia_1	0.4033

Table 8: Results of the HaSpeeDe-TW task.

HaSpeeDe-FB and HaSpeeDe-TW, i.e. ItaliaNLP, also achieved valuable results in the cross-domain sub-tasks, ranking at fifth and first position in Cross-HaSpeeDe_FB and Cross-HaSpeeDe_TW, respectively. But these results can also depend on the association of the polarity and subjectivity lexicon with word embeddings, which alone did not allow the achievement of particularly high results.

Furthermore, it is not surprising that the best results have been obtained on HaSpeeDe-FB, provided the fact that messages posted on this platform are longer and more correct than those in Twitter, allowing systems (and humans too) to find more and more clear indications of the presence of HS.

The coarse granularity of the annotation scheme,

Team	Macro F1-score
baseline	0.4033
InriaFBK_2	0.6541
InriaFBK_1	0.6531
VulpeculaTeam	0.6542
Perugia_2	0.6279
ItaliaNLP_1	0.6068
ItaliaNLP_2	0.5848
GRCP_2	0.5436
RuG_1	0.5409
RuG_2	0.4845
GRCP_1	0.4544
HanSEL	0.4502
StopPropagHate	0.443
Perugia_1	0.4033

Table 9: Results of the Cross-HaSpeeDe_FB sub-task.

Team	Macro F1-score
baseline	0.2441
ItaliaNLP_2	0.6985
InriaFBK_2	0.6802
ItaliaNLP_1	0.6693
InriaFBK_1	0.6547
VulpeculaTeam*	0.6189
RuG_1	0.6021
RuG_2	0.5545
HanSEL	0.4838
Perugia_2	0.4594
GRCP_1	0.4451
StopPropagHate*	0.4378
GRCP_2	0.318
Perugia_1	0.2441

Table 10: Results of the Cross-HaSpeeDe_TW sub-task.

which is a simplification of the schemes originally proposed for the datasets, and merged specifically for the purpose of this task, probably influenced the scores which are indeed very promising and high with respect to other tasks of the sentiment analysis area.

As regards the Cross-HaSpeeDe_FB and Cross-HaSpeeDe_TW sub-tasks, the lower results with respect to the in-domain tasks can be attributed to several factors, among which - and as expected - the different distribution in Facebook and Twitter datasets of HS and not HS classes. As a matter of fact, the percentage of HS in the Facebook train and test set is around 46% and 68%, respectively, while in the Twitter test set is around 32% in both sets. Such imbalanced distribution is reflected in the overall system outputs in the two sub-tasks: in Cross-HaSpeeDe_FB, where systems have been evaluated against the Twitter test set, most of the labels predicted as HS were not classified as such in the gold standard; conversely, in Cross-HaSpeeDe_TW, the majority of labels predicted as not HS were actually considered as HS in the gold corpus.
Another feature that distinguishes Facebook from Twitter dataset is the wider range of hate categories in the former, compared to the latter (see Section 3.1 and 3.2). Especially in Cross-HaSpeeDe_TW, the identification of hateful messsages may have been made even more difficult due to the reduced number of potential hate targets in the training set, with respect to the test set.

Overall, the heterogeneous nature of the datasets provided for the task - both in terms of class distribution and data composition - together with their quite small size, made the whole task even more challenging; nonetheless, this did not prevent participants from finding the appropriate solutions, thus improving the state of the art for HS identification in Italian language as well.

6 Closing Remarks

The paper describes the HaSpeeDe task for the detection of HS in Italian texts from Facebook and Twitter. The novelty of the task mainly consists in allowing the comparison between the results obtained on the two platforms and experiments on training on one typology of texts and testing on the other. The results confirmed the difficulty of cross-platform HS detection but also produced very promising scores in the tasks where the data from the same social network were exploited both for training and testing.
Future work can be devoted to an in-depth analysis of errors and to the observation of the contribution that different resources can give to systems performing this task.

Acknowledgments

The work of Cristina Bosco and Manuela Sanguinetti is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

References

Xiaoyu Bai, Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. RuG @ EVALITA 2018: Hate Speech Detection In Italian Social Media. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*.

Giulio Bianchini, Lorenzo Ferri, and Tommaso Giorni. 2018. Text Analysis for Hate Speech Detection in Italian Messages on Twitter and Facebook. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Cristina Bosco, Patti Viviana, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Ruffo, Rossano Schifanella, and Marco Stranisci. 2017. Tools and Resources for Detecting Hate and Prejudice Against Immigrants in Social Media. In *Proceedings of First Symposium on Social Interactions in Complex Intelligent Systems (SICIS), AISB Convention 2017, AI and Society*.

Pete Burnap and Matthew L. Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, 7(2).

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *IberEval@SEPLN*. CEUR-WS.org.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Irony Detection in Italian Tweets (IronITA). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task Learning in Deep Neural Networks at EVALITA 2018. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing Different Supervised Approaches to Hate Speech Detection. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *CoRR*, abs/1703.04009.

Gretel Liz De la Peña Sarracén, Reynaldo Gil Pons, Carlos Enrique Muñiz Cuza, and Paolo Rosso. 2018. Hate Speech Detection Using Attention-based LSTM. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*.

Karmen Erjavec and Melita Poler Kovačič. 2012. "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society*, 15(6).

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@SEPLN*. CEUR-WS.org.

Paula Fortuna, Ilaria Bonavita, and Sérgio Nunes. 2018. Merging datasets for hate speech classification in Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. In *Proceedings of the First Workshop on Abusive Language*.

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).

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics.

Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets Against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press.

Yashar Mehdad and Joel Tetreault. 2016. Do Characters Abuse More Than Words? In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

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, UMAP 2016*.

Serena Pelosi, Alessandro Maisto, Pierluigi Vitale, and Simonetta Vietri. 2017. Mining Offensive Language on Social Media. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR.

Marco Polignano and Pierpaolo Basile. 2018. HanSEL: Italian Hate Speech Detection through Ensemble Learning and Deep Neural Networks. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

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 Language Resources and Evaluation Conference 2018*.

Valentino Santucci, Stefania Spina, Alfredo Milani, Giulio Biondi, and Gabriele Di Bari. 2018. Detecting Hate Speech for Italian Language in Social Media. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

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

Dirk von Grünigen, Ralf Grubenmann, Fernando Benites, Pius Von Däniken, and Mark Cieliebak. 2018. spMMMP at GermEval 2018 Shared Task: Classification of Offensive Content in Tweets using Convolutional Neural Networks and Gated Recurrent Units. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017. *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.