HaSpeeDe 2 @ EVALITA2020

Sanguinetti, Manuela; Comandini, Gloria; di Nuovo, Elisa; Frenda, Simona; Stranisci, Marco; Bosco, Cristina; Caselli, Tommaso; Patti, Viviana; Russo, Irene

Published in:
Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)

HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task

Manuela Sanguinetti[⋆], **Gloria Comandini**[⋄], **Elisa di Nuovo**[⊕], **Simona Frenda**[⊕], **Marco Stranisci**[⊕], **Cristina Bosco**[⊕], **Tommaso Caselli**[⊖], **Viviana Patti**[⊕], **Irene Russo**[†]
[⋆]Università degli Studi di Cagliari, [⋄] Università degli Studi di Trento,
[⊕]Università degli Studi di Torino, [⊖]University of Groningen
[†]ILC-CNR Pisa
manuela.sanguinetti@unica.it, gloria.comandini@unitn.it,
{dinuovo,frenda,stranisc,bosco,patti}@di.unito.it,
t.caselli@rug.nl, irene.russo@ilc.cnr.it

Abstract

The Hate Speech Detection (HaSpeeDe 2) task is the second edition of a shared task on the detection of hateful content in Italian Twitter messages. HaSpeeDe 2 is composed of a Main task (hate speech detection) and two Pilot tasks, (stereotype and nominal utterance detection). Systems were challenged along two dimensions: (i) time, with test data coming from a different time period than the training data, and (ii) domain, with test data coming from the news domain (i.e., news headlines). Overall, 14 teams participated in the Main task, the best systems achieved a macro F1-score of 0.8088 and 0.7744 on the in-domain in the out-of-domain test sets, respectively; 6 teams submitted their results for Pilot task 1 (stereotype detection), the best systems achieved a macro F1-score of 0.7719 and 0.7203 on in-domain and out-of-domain test sets. We did not receive any submission for Pilot task 2.

1 Introduction and Motivations

From a NLP perspective, much attention has been paid to the automatic detection of Hate Speech (HS) and related phenomena (e.g., offensive or abusive language among others) and behaviors (e.g., harassment and cyberbullying). This has led to the recent proliferation of contributions on this topic (Nobata et al., 2016; Waseem et al., 2017; Fortuna et al., 2019), corpora and lexica[1], dedicated workshops[2], and shared tasks within national[3] and international[4] evaluation campaigns.

As for Italian, the first edition of HaSpeeDe (Bosco et al., 2018), a task specifically focused on HS detection, was proposed at EVALITA 2018 (Caselli et al., 2018). The task consisted of the binary classification (HS vs not-HS) of texts from Twitter and Facebook. For each social media platform, training and test data were provided. Furthermore, two cross-platform sub-tasks were introduced to test the systems' ability to generalize across platforms.

The ultimate goal of HaSpeeDe 2 at EVALITA 2020 (Basile et al., 2020) is to take a step further in state-of-the-art HS detection for Italian. By doing this, we also intend to explore other side phenomena and see the extent to which they can be automatically distinguished from HS.

We propose a single training set made of tweets, but two separate test sets within two different domains: tweets and news headlines. While social media are still one of the main channels used to spread hateful content online (Alkiviadou, 2019; Wodak, 2018), an important role in this respect is also played by traditional media, and newspapers in particular.

Furthermore, we chose to include another HS-related phenomenon, namely the presence of stereotypes referring to one of the targets identified within our dataset (i.e., muslims, Roma and immigrants). With the term stereotype we mean any explicit or implicit reference to typical beliefs and attitudes about a given target (Sanguinetti et al., 2018). An error analysis of the main systems on the HaSpeeDe 2018 dataset itself (Francesconi

[1]More details and an overview of available HS resources have been recently presented in Poletto et al. (2020).

[2]More detailed informations in: https://www.workshopononlineabuse.com/

[3]HASOC (Mandl et al., 2019), Poleval (Ptaszynski et al., 2019) or VLSD (Vu et al., 2019).

[4]Hateval task at Semeval 2019 (Basile et al., 2019).

et al., 2019) showed that the occurrence of these elements constitutes a common source of error in HS identification.

Finally, it has been observed that in social media and newspapers' headlines, the most hateful parts are often verbless sentences or a verbless fragments, also known as Nominal Utterances (NUs) (Comandini et al., 2018). The relevant presence of NUs has been investigated in the POP-HS-IT corpus (Comandini and Patti, 2019). In order to have a better understanding of the syntactic strategies used in HS, we include the recognition of NUs in hateful tweets and news headlines.

2 Task Description

HaSpeeDe 2[5] consists of a Main task and two Pilot tasks and is based on two datasets, one containing messages from a social media platform, namely Twitter, and the other one news headlines. The three tasks are shortly described as follows:

- **Task A - Hate Speech Detection (Main Task):** binary classification task aimed at determining the presence or the absence of hateful content in the text towards a given target (among immigrants, Muslims and Roma)
- **Task B - Stereotype Detection (Pilot Task 1):** binary classification task aimed at determining the presence or the absence of a stereotype towards the same targets as Task A
- **Task C - Identification of Nominal Utterances (Pilot Task 2):** sequence labeling task aimed at recognizing NUs in data previously labeled as hateful.

This edition of the task presents several distinguishing features with respect to the first one. Besides including new and more-richly annotated data, news headlines were introduced as cross-domain test data. Furthermore, two additional tasks are proposed. Finally, the Twitter test set intentionally contains tweets published in a different time frame than those in the training set to verify the systems' ability to detect HS forms independently of biases. These biases result from context-related features, such as events – regarding one of our HS targets – that can be controversial or be subject to heated and polarized debates.

[5]Task repository: https://github.com/msang/haspeede/tree/master/2020.

3 Datasets and Formats

In this section we describe the datasets and formats used in the three tasks.

3.1 Twitter Dataset

Task A: The Twitter portion of the data of HaSpeeDe 2018 was included in the training set (4,000 tweets posted from October 2016 to April 2017). Moreover, new Twitter data were included for this competition, a subset of the data gathered for the Italian hate speech monitoring project "Contro l'Odio" (Capozzi et al., 2019). The data were retrieved using the Twitter Stream API and filtered using the set of keywords described in Poletto et al. (2017). The newly annotated tweets were posted between September 2018 and May 2019 and were annotated by Figure Eight (now Appen) contributors for hate speech and by the task organizers for the stereotype category. In particular, only data posted between January and May 2019 were included in the test set.

Task B: The HaSpeeDe Twitter corpus – used in the first edition of the task – was already annotated for stereotype since it was part of the Italian Hate Speech corpus described in Sanguinetti et al. (2018). We then used the same guidelines to enrich the new data from "Contro l'Odio" with this annotation layer. The annotation was carried out by the task organizers.

Task C: The HaSpeeDe Twitter corpus was also annotated for the presence of Nominal Utterances (NUs) within a side project (Comandini and Patti, 2019). We used an updated version of its guidelines (available in the task repository) to enrich the new hateful data introduced in the campaign. Similarly to the stereotype level, the annotation of NUs was carried out by the task organizers specifically for this task's purposes.

3.2 News Dataset

Task A: For task A a new test corpus composed of newspapers' headlines about immigrants was made available. The data were retrieved between October 2017 and February 2018 from online newspapers (*La Stampa*, *La Repubblica*, *Il Giornale*, *Liberoquotidiano*) and annotated within the context of a Master's degree thesis discussed in 2018 at the Department of Foreign Languages at the University of Turin. Data annotation includes

the same categories annotated in the Twitter corpus.

Task B: The News corpus also includes stereotype annotation, performed according to the same guidelines used for developing the Twitter corpus.

Task C: Similarly to the Twitter dataset, the third annotation level was added in the News corpus from scratch and specifically for the present task.

Tables 1, 2 and 3 show the data distribution for each task.

TASK A	HS	NOT HS	TOT.
Train	2766	4073	6839
Test Tweets	622	641	1263
Test News	181	319	500

Table 1: Distribution of Hate Speech labels.

TASK B	STER.	NOT STER.	TOT.
Train	3042	3797	6839
Test Tweets	569	694	1263
Test News	175	325	500

Table 2: Distribution of Stereotype labels.

TASK C	w/ NUs	w/o NUs	TOT.
Train	1565	1201	2766
Test Tweets	379	243	622
Test News	151	30	181

Table 3: Distribution of Nominal Utterances.

The whole dataset consists of 8,012 tweets and 500 news headlines for Task A and B, and 3,388 tweets and 181 news (i.e., the sub-set with hateful data only) for Task C.
In Task A and B, HS and stereotype represent the 41.8% and 44.6%, respectively, of the Twitter dataset. In contrast, in the News dataset, the portion of hateful content and stereotype lowers to 36% and 35%.
Table 3 shows statistics about the total number of texts with or without NUs in Task C. The percentage of hateful tweets featuring at least one NU is 57.4%; the percentage of news headlines having at least one NU is 83.4%. This distribution is in line with the one found in Comandini and Patti (2019).

3.3 Formats

Task A and B: For both tasks A and B data are provided in a tab-separated values (TSV) file including ID, text, HS and stereotype class (`0` or `1`). Mentions and URLs were replaced with `@user` and `URL` placeholders. Table 4 shows some annotation examples.

Task C: The dataset provided for Task C was annotated using WebAnno and converted into a IOB (Inside-Outside-Beginning) format. The resulting IOB2 alphabet consists of I-NU-CGA, O and B-NU-CGA.
The annotation includes the ID, followed by an hyphen to mark the token number, the token, and the IOB2 annotation of the NUs.
Below an example taken from the training set.

#Text=È UNA PROVOCAZIONE...ORA BASTA.. NESSUNO SBARCHI IN #ITALIA[6]

9602-23	È	O
9602-24	UNA	O
9602-25	PROVOCAZIONE	O
9602-26	.	O
9602-27	.	O
9602-28	.	O
9602-29	ORA	B-NU-CGA
9602-30	BASTA	I-NU-CGA
9602-31	.	I-NU-CGA
9602-32	.	I-NU-CGA
9602-33	NESSUNO	O
9602-34	SBARCHI	O
9602-35	IN	O
9602-36	#	O
9602-37	ITALIA	O

To prevent participants from cheating, the released test set for Task C also contains non-hateful messages. However, the evaluation of the systems is conducted only on the hateful messages since we are interested in investigating the relationship between these two phenomena.

4 Evaluation

For each task, participants were allowed to submit up to 2 runs. A separate official ranking was provided, and the evaluation was performed according to the standard metrics, i.e, Precision, Recall and F-score.
For Task A and Task B, the scores were computed for each class separately, and finally the F-score was macro-averaged to get the overall results.

[6] "IT'S A PROVOCATION...THAT'S ENOUGH...NO LANDINGS IN #ITALY"

id	text	hs	ster.
8783^T	*Via tutti i campi Rom e disinfettare per bene il lerciume che si lasciano dietro.* *Mai più campi Rom in Italia NO NO E NO* ("Away all the Roma camps and clean the filth they leave behind. No more Roma camps in Italy NO NO AND NO")	1	1
9254^T	*Vanno affondate. Hanno rotto i c.....i* *Aquarius vuol dettare ancora legge: carica migranti e rifiuta gli ordini libici* ("They must be sunk. We've had enough Aquarius still wants to lay down the law: it brings migrants on board and refuses Lybian orders")	1	0
9414^T	*Istat conferma: migranti vengono in Italia a farsi mantenere* ("Istat confirms: migrants come to Italy to sponge off (us)")	0	1
10707^N	*Sea Watch, Finanza sequestra la nave: sbarcano i migranti* ("Sea Watch, Custom Corps confiscate the ship: migrants get off")	0	0

Table 4: Examples from the datasets for Task A and B. ^{T} and ^{N} superscripts indicate, respectively, whether the message is from the Twitter or News dataset.

For Task C, token-wise scores were computed, and a NU was considered correct only in case of exact match, i.e., if all tokens that compose it were correctly identified.

Different baseline systems were built according to the task type:

- For Task A and B, besides a typical classifier based on the most frequent class (Baseline_MFC in Tables 5–8), a Linear SVM with TF-IDF of unigrams and 2–5 char-grams was used (Baseline_SVC).
- For Task C, the baseline replicates the one presented for the COSMIANU corpus (Comandini et al., 2018), which identifies as correct in the test the NUs that appear in the training set (memory-based approach); baseline results in Table 9.

5 Task Overview: Participation and Results

5.1 Participants

A total amount of 14 teams participated in the Main task on HS detection, 6 teams also submitted their results for the Pilot task 1 (i.e. Task B) on stereotype detection, while we did not receive any submission for the Pilot task 2 (i.e. Task C) on NUs identification. Except for one case, all teams submitted 2 runs for their tasks. Furthermore, 4 teams used the same systems to participate in other (and partly related) tasks within the EVALITA 2020 campaign: YNU_OXZ and Jigsaw participated in the task on Automatic Misogyny Identification (AMI) (Fersini et al., 2020), while TextWiller and Venses also participated in the task on Stance Detection in Italian Tweets (SardiStance) (Cignarella et al., 2020). It is worth pointing out that in this second edition we registered a higher participation of non-Italian and non-academic teams, and that HaSpeeDe 2 has been one of the most participated EVALITA 2020 tasks.

5.2 Systems Overview

Approaches The participating models are characterized by different architectures that exploit principally BERT-based models and linguistic features. Transformers are a popular choice in this edition. Jigsaw (Lees et al., 2020), Svandiela (Klaus et al., 2020), DH-FBK (Leonardelli et al., 2020), TheNorth (Lavergne et al., 2020) fine-tuned BERT, AlBERTo[7] and UmBERTo[8] language models for both runs. YNU_OXZ (Ou and Li, 2020) exploited the pre-trained XLM-RoBERTa[9] multi-language model as input of Neural Networks architecture. Fontana-Unipi (Fontana and Attardi, 2020) developed a model that is an ensemble of fixed number of instances of two principal transformers (AlBERTo and DBMDZ[10]) and a combination of DBMDZ input and a dense layer. The DBMDZ is used also by By1510 (Deng et al., 2020) in a transfer learning approach. UO team (Rodriguez Cisnero and Ortega Bueno, 2020), on the other hand, used a Bi-LSTM with the addition of linguistic features in

[7] https://github.com/marcopoli/AlBERTo-it

[8] https://github.com/musixmatchresearch/umberto

[9] https://huggingface.co/transformers/model_doc/xlmroberta.html

[10] https://huggingface.co/dbmdz/bert-base-italian-uncased

the first run, while using the pre-trained DBMDZ model in the second one. CHILab (Gambino and Pirrone, 2020) experimented transformer encoders in the first run and depth-wise Separable Convolution techniques in the second one. Moreover, some teams explored classical machine learning approaches such as No Place For Hate Speech (dos S. R. da Silva and T. Roman, 2020), TextWiller (Ferraccioli et al., 2020), UR_NLP (Hoffmann and Kruschwitz, 2020) and Montanti (Bisconti and Montagnani, 2020). Finally, Venses (Delmonte, 2020), based on the parser for Italian ItGetaruns, applied six different rule-based classifiers.

Features and Lexical Resources Various features are tested and explored by participants. Morphosyntactic features are exploited by CHILab, using Part-of-Speech tags as additional input. To adapt the POS tagging model provided by Python's spaCy library to social media language, they added emoticons, emojis, hashtags and URLs to vocabulary. In addition, to preprocess the texts, they used sentiment lexicon for replacing emoticons with appropriate labels about the expressed sentiment. Semantic and lexical features are exploited by Venses and UO teams. In particular, UO team used WordNet to catch lexical ambiguity, syntactic patterns and similarity among words; calculated information gain to capture the most relevant words; used lexicons such as HurtLex (Bassignana et al., 2018) and SenticNet[11] to feature words with hateful categories and sentiment information. Finally, different types of representation of tweets are tested by Montanti: TF-IDF, DistilBert[12] and GloVe (Pennington et al., 2014) vectors as well as their combination.

Additional data Some teams preferred to use additional data to improve the knowledge of their classifiers. To extend the provided training set, YNU_OXZ exploited Facebook data provided in the first edition of HaSpeeDe and DH-FBK used a set of Italian tweets that covers similar topics. Jigsaw, for one of the submissions, used additional user-generated comments to fine-tune their model. CHILab used additional tweets taken from TWITA 2018[13] by means of some keywords extracted from the provided training set to extend the embedding layer of their model. Finally, the SENTIPOLC 2016 dataset was exploited by UO team.

Interaction between Task A and B Except for TheNorth team, most of the participants did not consider the interaction between Task A and B. Taking into account the possible correlation between texts containing hate speech and texts expressing stereotyped ideas about targets, TheNorth tested the performance of multitasking approach for both tasks (second run) against a fine-tuned UmBERTo model (first run). In particular, observing competition results we can notice the efficacy of multitasking in hate speech identification and not in stereotype detection.

5.3 Results

In Table 5, 6, 7 and 8, we report the official results of HaSpeeDe 2 for Task A and B, ranked by the macro-F1 score. In case of multiple runs, a suffix has been appended to each team name, in order to distinguish the run ID of the submitted file.

Team	Macro-F1
TheNorth_2	0.8088
TheNorth_1	0.7897
CHILab_1	0.7893
Fontana-Unipi	0.7803
CHILab_2	0.7782
By1510_1	0.7766
Svandiela_2	0.7756
YNU_OXZ_1	0.7717
Jigsaw_al	0.7681
UR_NLP_2	0.7598
DHFBK_2	0.7534
DHFBK_1	0.7495
No Place For Hate Speech_STT	0.7491
Svandiela_1	0.7452
Montanti_1	0.7432
UR_NLP_1	0.7399
YNU_OXZ_2	0.7345
Montanti_2	0.7279
UO_2	0.7214
Baseline_SVC	**0.7212**
Jigsaw_js	0.717
By1510_2	0.7065
No Place For Hate Speech_LRT	0.7057
UO_1	0.6878
Venses_1	0.5054
Venses_2	0.4726
TextWiller_1	0.3604
Baseline_MFC	**0.3366**
TextWiller_2	0.3317

Table 5: Task A results on Twitter data.

As a general remark, we can observe that the in-domain Main task registered better results (macro-F1=0.8088) both compared to the cross-domain counter-part (0.7744) and the Pilot task 1; in turn,

[11] https://www.sentic.net/

[12] https://huggingface.co/transformers/model_doc/distilbert.html

[13] http://twita.di.unito.it/

Team	Macro-F1
CHILab_1	0.7744
UO_2	0.7314
Montanti_1	0.7256
CHILab_2	0.7183
DHFBK_2	0.702
UR_NLP_2	0.6983
YNU_OXZ_2	0.6922
Montanti_2	0.6821
Jigsaw_js	0.6755
DHFBK_1	0.6744
TheNorth_1	0.671
UR_NLP_1	0.6684
UO_1	0.6657
By1510_2	0.6638
YNU_OXZ_1	0.6604
TheNorth_2	0.6602
Fontana-Unipi	0.6546
Jigsaw_al	0.6353
No Place For Hate Speech_STN	0.6328
No Place For Hate Speech_LRN	0.6212
Baseline_SVC	**0.621**
By1510_1	0.6094
Svandiela_2	0.6031
Svandiela_1	0.5265
Venses_1	0.5024
Baseline_MFC	**0.3894**
Venses_2	0.3805
TextWiller_1	0.3101
TextWiller_2	0.2693

Table 6: Task A results on News data.

better results were obtained in the latter with the in-domain data compared to the News set (0.7744 and 0.7203, respectively). The best performances overall provided by the systems used for Task A on Twitter data is also reflected in the average value of the macro-F1 scores of each ranking: 0.6899 for the latter, 0.6306 for Task B on Twitter data, 0.6144 for Task A on News data and 0.5972 for Task B on News data.

We also considered the overall results achieved by all participating teams and observed that, as regards Task A, 12 and 13 teams (in the Twitter and News test set, respectively) obtained higher scores than the SVM-based baseline with at least one of the submitted runs, and 13 teams, on both domains, outperformed the one based on the most frequent class. For Task B, and with respect to the SVM baseline, the same is true for 4 teams out of 6 in the Twitter set and for 3 teams in the News set, while all teams beat the majority-class baseline with at least one run.

Regarding Task C, since the training set is composed of tweets, we first investigated the macro F-score value on a validation set created by splitting the training set in 80%-20%. We then tested the memory-based baseline described in Section

Team	Macro-F1
TheNorth_1	0.7719
TheNorth_2	0.7676
CHILab_1	0.7615
Jigsaw_al	0.7415
CHILab_2	0.7386
Baseline_SVC	**0.7149**
Montanti_1	0.7076
Montanti_2	0.6889
Jigsaw_js	0.6674
TextWiller_2	0.6031
Venses_1	0.5078
Venses_2	0.4671
Baseline_MFC	**0.3546**
TextWiller_1	0.3369

Table 7: Task B results on Twitter data.

Team	Macro-F1
CHILab_1	0.7203
CHILab_2	0.7184
Montanti_1	0.7166
TheNorth_1	0.6854
Jigsaw_al	0.6811
Montanti_2	0.6706
Baseline_SVC	**0.6688**
TheNorth_2	0.6465
Jigsaw_js	0.6412
TextWiller_2	0.6053
Venses_1	0.5386
Baseline_MFC	**0.3939**
Venses_2	0.3671
TextWiller_1	0.3077

Table 8: Task B results on News data.

4 on the two test sets released for the task. Table 9 shows the macro-F1 values obtained in the validation set, in the Twitter test set as well as in the News test set. As mentioned earlier, no submissions were made for this task, but the baselines' values for both domains are reported in this overview as reference points for further works.

Baseline	Macro-F
Baseline_validation	**0.1459**
Baseline_test_Tweets	0.0706
Baseline_test_News	0.0087

Table 9: Task C - Baseline results for Tweets and News.

6 Discussion

A discussion of results, especially those regarding the Main task, necessarily involves a preliminary comparison with the ones obtained in the first edition of HaSpeeDe, in particular in the two tasks where Twitter data were used for training, i.e. HaSpeeDe_TW and Cross-HaSpeeDe_TW.

The best systems attained macro-F1=0.7993 in the former task and 0.6985 in the latter. While these results are in line with those reported for Task A on the in-domain data, the results obtained in this edition on News data are better than the part cross-domain task, where the test set was made up of Facebook comments. We hypothesize that the homogeneity of hate target in News and Twitter corpora (immigrants) has meant more than the similar linguistic features in Twitter and Facebook data, stemming from the fact that they are both social media texts.

Participants achieved promising results in the detection of stereotypes, a new pilot task proposed at HaSpeeDe this year for the first time. In our view, stereotype and HS are meant as orthogonal dimensions of abusive language, which do not necessarily coexist. This influenced the design of HaSpeeDe 2, where we proposed two independent tasks for the detection of such categories. However, a first analysis of systems participating in both tasks suggests that most teams did not design a dedicated system for stereotype recognition, but focused on developing a HS detection model, adapting the same model to stereotype recognition, reducing *de facto* stereotypes to characteristics of HS. We hypothesize that this could be one of the factors that led the systems to not generalize well when applied to the stereotype detection task, especially in texts that are not hateful but contain stereotypes. This hypothesis is confirmed by the high percentage of false negatives (21% in tweets and 35% in news headlines) of the stereotype class in non-hateful texts, with respect to false negatives (5% in tweets and 28% in news headlines) in hateful ones. It is possible to notice the same increase also in false positives in hateful texts. These values suggest that stereotype appears as a more subtle phenomenon that could not give rise to hurtful message. The percentages have been computed taking into account the set of common incorrect predictions of the three best runs in Task B, and calculated in relation to the actual distribution of HS and stereotype in the test set. Analyzing the predictions of the three best runs in Task A, similar influence of stereotype is observed in false negative and positive, but to a minor extent. These results are in line with the observations about emerged from the error analysis of HaSpeeDe 2018 (Francesconi et al., 2019).

To conclude the discussion on this edition's results, we comment on the baseline scores obtained for Task C. As it can be noticed from Table 9, the value obtained on the validation set is higher than the ones obtained on both test sets. This variation can be explained by the main characteristics of the data at hand: on the Twitter side, this is due to the different time frames of tweet's publication included in training and test set, while on the News side, such low value is expected by virtue of the different text domain. Since this baseline uses a memory-based approach, such a low performance is to be expected in datasets from different time frames, since the discussion topics are different and Twitter users change their hashtags and slogans, which are the main repeated items.

7 Conclusions

In its second edition, the HaSpeeDe task proposed the detection of hateful content in Italian, by challenging systems along two dimensions, time and domain, and taking into account also the category of stereotype, which often co-occurs with HS. This paves the way for further investigations also about the relationships linking stereotype and HS.

In order to take a step further in state-of-the-art HS detection, the task provided novel benchmarks for exploring different facets of the phenomenon and laying the foundations for deeper studies about the impact of bias, topic and text domain. In this line, also a pilot task about recognition of NUs was proposed, devoted to study this kind of linguistic form in hateful messages in tweets and newspaper headlines, as it has been proved that both headlines in journalistic writings (Mortara Garavelli, 1971) and social media texts (Ferrari, 2011; Comandini et al., 2018) are a fertile ground for NUs. Even though we did not receive any submission for Pilot task 2, our hope is that the fine-grained annotation of hateful data concerning these aspects can be the subject of deeper studies to shed light on the syntax of hate, a topic still understudied.

Acknowledgments

The work of Cristina Bosco, Simona Frenda, Viviana Patti and Marco Stranisci is partially funded by Progetto di Ateneo/CSP 2016 (Immigrants, Hate and Prejudice in Social Media, S1618.L2.BOSC.01) and by the project "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call).

References

Natalie Alkiviadou. 2019. Hate speech on social media networks: towards a regulatory framework? *Information & Communications Technology Law*, 28(1):19–35.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of SemEval 2019*.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online.

Elisa Bassignana, 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)*.

Elia Bisconti and Matteo Montagnani. 2020. Montanti @ HaSpeeDe2 EVALITA 2020: Hate Speech Detection in online contents. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, Giovanni Semeraro, and Marco Stranisci. 2019. Computational linguistics against hate: Hate speech detection and visualization on social media in the "Contro L'Odio" project. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, CLiC-it 2019*.

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)*.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Gloria Comandini 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*.

Gloria Comandini, Manuela Speranza, and Bernardo Magnini. 2018. Effective Communication without Verbs? Sure! Identification of Nominal Utterances in Italian Social Media Texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253. CEUR-WS.org.

Rodolfo Delmonte. 2020. Venses @ HaSpeeDe2 & SardiStance: Multilevel Deep Linguistically Based Supervised Approach to Classification. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Tao Deng, Yang Bai, and Hongbing Dai. 2020. By1510 @ HaSpeeDe 2: Identification of Hate Speech for Italian Language in Social Media Data. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Adriano dos S. R. da Silva and Norton T. Roman. 2020. No Place For Hate Speech @ HaSpeeDe 2: Ensemble to identify hate speech in Italian. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Federico Ferraccioli, Andrea Sciandra, Mattia Da Pont, Paolo Girardi, Dario Solari, and Livio Finos. 2020. TextWiller @ SardiStance, HaSpeede2: Text or Con-text? A smart use of social network data in predicting polarization. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Angela Ferrari. 2011. Enunciati nominali. *Enciclopedia dell'Italiano*. `http://www.treccani.it/enciclopedia/enunciati-nominali_(Enciclopedia_dell'Italiano)/`.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic Misogyny Identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online.

Michele Fontana and Giuseppe Attardi. 2020. Fontana-Unipi @ HaSpeeDe2: Ensemble of transformers for the Hate Speech task at Evalita. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the Third Workshop on Abusive Language Online*.

Chiara Francesconi, Cristina Bosco, Fabio Poletto, and Manuela Sanguinetti. 2019. Error Analysis in a Hate Speech Detection Task: The case of HaSpeeDe-TW at EVALITA 2018. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*.

Giuseppe Gambino and Roberto Pirrone. 2020. CHILab @ HaSpeeDe 2: Enhancing Hate Speech Detection with Part-of-Speech Tagging. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Julia Hoffmann and Udo Kruschwitz. 2020. UR_NLP @ HaSpeeDe 2 at EVALITA 2020: Towards Robust Hate Speech Detection with Contextual Embeddings. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Svea Klaus, Anna-Sophie Bartle, and Daniela Rossmann. 2020. Svandiela @ HaSpeeDe: Detecting Hate Speech in Italian Twitter Data with BERT. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Eric Lavergne, Rajkumar Saini, György Kovács, and Killian Murphy. 2020. TheNorth @ HaSpeeDe 2: BERT-based Language Model Fine-tuning for Italian Hate Speech Detection. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Elisa Leonardelli, Stefano Menini, and Sara Tonelli. 2020. DH-FBK @ HaSpeeDe2: Italian Hate Speech Detection via Self-Training and Oversampling. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*.

Bice Mortara Garavelli. 1971. Fra norma e invenzione: lo stile nominale. In Accademia della Crusca, editor, *Studi di grammatica italiana*, volume 1, pages 271–315. G. C. Sansoni Editore, Firenze, Italia.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*.

Xiaozhi Ou and Hongling Li. 2020. YNU_OXZ @ HaSpeeDe 2 and AMI : XLM-RoBERTa with Ordered Neurons LSTM for classification task at EVALITA 2020. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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)*.

Fabio Poletto, 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*.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter. In *Proceedings of the PolEval 2019 Workshop*.

Mariano Jason Rodriguez Cisnero and Reynier Ortega Bueno. 2020. UO@HaSpeeDe2: Ensemble Model for Italian Hate Speech Detection. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

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 Eleventh International Conference on Language Resources and Evaluation (LREC'18)*.

Xuan-Son Vu, Thanh Vu, Mai-Vu Tran, Thanh Le-Cong, and Huyen T M Nguyen. 2019. HSD Shared Task in VLSP Campaign 2019: Hate Speech Detection for Social Good. In *Proceedings of VLSP 2019*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Ruth E. Wodak. 2018. Introductory remarks from 'hate speech' to 'hate tweets'. In Mojca Pajnik and Birgit Sauer, editors, *Populism and the web: communicative practices of parties and movements in Europe*, pages xvii–xxiii. Rourledge.