DH-FBK @ HaSpeeDe2: Italian Hate Speech Detection via Self-Training and Oversampling

Elisa Leonardelli
Fondazione Bruno Kessler
Trento, Italy
eleonardelli@fbk.eu

Stefano Menini
Fondazione Bruno Kessler
Trento, Italy
menini@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Abstract

We describe in this paper the system submitted by the DH-FBK team to the HaSpeeDe evaluation task, and dealing with Italian hate speech detection (Task A). While we adopt a standard approach for fine-tuning AlBERTo, the Italian BERT model trained on tweets, we propose to improve the final classification performance by two additional steps, i.e. self-training and oversampling. Indeed, we extend the initial training data with additional silver data, carefully sampled from domain-specific tweets and obtained after first training our system only with the task training data. Then, we retrain the classifier by merging silver and task training data but oversampling the latter, so that the obtained model is more robust to possible inconsistencies in the silver data. With this configuration, we obtain a macro-averaged F1 of 0.753 on tweets, and 0.702 on news headlines.

1 Introduction

Although hate speech detection may seem a solved task on English, with more than 60 systems participating in the last Offenseval edition reaching an F1 $>$ 0.90 (Zampieri et al., 2020), this goal has not been reached when moving to other languages and settings. For example, at the last HaSpeeDe shared task on Italian (Bosco et al., 2018) the best systems reached 0.83 F1 on Facebook data and 0.80 on Twitter data (Cimino et al., 2018), but the performance dropped below 0.70 F1 when dealing with a cross-domain setting, i.e. training on Facebook and testing on Twitter (Cimino et al., 2018), and vice-versa (Corazza et al., 2018). Other recent studies confirmed that detecting hate speech on different social media platforms would require a platform-specific setting, and that just merging all training data coming from different sources does not always improve performance, in particular when testing on Twitter (Corazza et al., 2019).

The problem of developing hate speech detection systems that are robust when analysing different sources or data that vary over time is however an understudied problem. Therefore, the task of out-of-domain classification introduced this year at HaSpeeDe is particularly important and will hopefully foster the development and evaluation of classifiers with good generalisation capabilities.

Concerning our classification approach, we build a standard pipeline based on AlBERTo (Polignano et al., 2019b), the Italian transformer-based model trained on Twitter data, since BERT-like models represent the state of the art for hate speech detection (Zampieri et al., 2020). We extend it in two ways: first, we use *self-training* to build a first classifier with the task training data and annotate a large set of tweets collected via Islam- and immigrant-specific hashtags. The silver data and the task training set are then merged to train a second, possibly more robust classifier, which we use to classify the test set. When retraining, we introduce *over-sampling* in one of the two runs submitted by our team, i.e. we repeat five times the task training data so that they are balanced with respect to the silver data. This, together with self-training, proved to be effective when evaluated in a five-fold fashion on the training set, outperforming a standard approach based only on fine-tuning with AlBERTo.

2 Related Work

While most approaches to hate speech detection have been proposed for English, other systems have been recently developed to deal with a num-

ber of other languages, including Turkish, Arabic, Danish (Zampieri et al., 2020), German (Wiegand et al., 2018) and Spanish (Basile et al., 2019). Concerning Italian, the first *Hate Speech Detection* task (HaSpeeDe) for Italian was organized at EVALITA-2018 (Bosco et al., 2018). The task consisted in automatically annotating messages from Twitter and Facebook, with a boolean value indicating the presence (or not) of hate speech. The participating systems adopt a wide range of approaches, including bi-LSTM (la Peña Sarracén et al., 2018), SVM (Santucci et al., 2018), ensemble classifiers (Polignano and Basile, 2018; Bai et al., 2018), RNN (Fortuna et al., 2018), CNN and GRU (von Grunigen et al., 2018). The authors of the best-performing system, ItaliaNLP (Cimino et al., 2018), experiment with 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). The same training and test set released for HaSpeeDe have been recently used also for other types of evaluation, for example to compare classifier performance and settings across different languages (Corazza et al., 2020), confirming the importance of domain-specific language models and the effectiveness of deep learning approaches (in this case, LSTM + fasttext embeddings). Since the development of BERT-like transformer-based models, however, they have become state-of-the-art approaches in several NLP tasks. This includes also hate speech detection for Italian, with the BERT model AlBERTo (Polignano et al., 2019b), which has recently achieved top-scores in two out of three HaSpeeDe 2018 tasks (Polignano et al., 2019a). For this reason, we decided to develop a classifier using the same model and the same approach.

3 Task Description

For the 2020 edition of EVALITA (Basile et al., 2020), the HaSpeeDe task (Sanguinetti et al., 2020) has focused on three main phenomena relevant to online hate speech detection by proposing three different tasks:

- Task A (main task): binary classification task aimed at determining whether a message contains hate speech or not
- Task B: binary classification task aimed at determining whether a message contains stereotypes or not
- Task C: sequence labeling task aimed at recognizing nominal utterances in hateful tweets

We participate in Task A, which in 2020 has the goal also to investigate variation in language and time concerning hate speech detection. To this purpose, the training set contains Twitter data, accompanied by a test set including both in-domain and out-of-domain data (tweets + news headlines), as well as from different time periods.

4 Data

In our experiments we use two types of data, the HaSpeeDe2 dataset provided by the task organisers, and domain-specific data collected from Twitter, that we include as silver data. The two datasets are described below.

4.1 HaSpeeDe2 Dataset

This dataset contains the training data provided by the organizers. These data specifically focus on the presence or the absence of hateful content towards immigrants, muslims or roma people. It consists of 6,839 annotated tweets, with 2,766 messages annotated as hateful and 4,073 as non-hateful.

4.2 Silver data description

Since the task is focused on hate speech against immigrants and minorities, we decided to exploit a set of tweets in Italian that covers similar topics and that was collected within the European project Hatemeter[1] (Ferret et al., 2019). For this project, conducted between February 2018 and January 2020, we downloaded tweets using hashtags of hate towards the Islam community, for example *#nomoschee*, *#stopIslam*, etc. Even if the dataset mainly covers Islam, references to other minorities like Roma or generic Immigrants are also present. To ensure that also other minorities are well represented, we randomly select from this dataset tweets that contain the most common words as chosen from the training data provided by task organizers, i.e. *Rom*, *nomade*, *migrante*, *straniero*, *profugo*, *islam*, *mussulmano* (*musulmano*), *terrorista*. Overall, around 20,400 additional tweets were selected. We then perform a first round of

[1] http://hatemeter.eu/

classification of the "new" tweets using the available data provided by organizers as training. This results in a new silver dataset composed of 11,129 hate and 9,254 non-hate tweets. This additional dataset is then merged with the task gold data and used to re-train the classifier. Details are reported in the following Section.

5 System Description

The classifier developed for both runs submitted by our team is based on the Italian BERT model trained on tweets, called AlBERTo (Polignano et al., 2019b). After fine-tuning it on the task training data, we use the obtained classifier to automatically annotate the additional dataset described in Section 4.2. These silver data are then merged with the task training data and used to fine-tune AlBERTo a second time. For one of the two submitted runs, we also experiment with oversampling as follows:

- **Run1:** we add the silver data to the tweets provided by the organizers for the training, keeping 500 of the released tweets for validation. In this setting, the training set size is ∼27,000 tweets, including 20,400 silver instances.

- **Run2:** we add the silver data to the tweets provided by the organizers as in Run1, but the tweet from organizers are oversampled by repeating them five times (and shuffling) in the training set, while tweets from the silver dataset occur only once. In this setting, the training set includes ∼52,000 tweets, with 39% of them being silver data.

We tested also the option to automatically assign a *tag* to each tweet, stressing the presence of a certain topic (immigrants/roma people/islam) using a keyword-based approach. However, with this additional information the classifier performed worse than without any topic indicator, so we removed it from the final runs. Below we report a detailed description of the process to select the best classification model, and of the preprocessing steps.

5.1 Model selection

The best performance in a wide variety of NLP tasks is currently obtained with approaches based on BERT (Devlin et al., 2019), a pre-trained transformed-based language model that can be fine-tuned and adapted to specific tasks by adding just one additional output layer to the neural network. As different BERT models exist, we first evaluated whether to use a multilingual version of BERT or the Italian version trained on Twitter data, called AlBERTo (Polignano et al., 2019b).

The comparison and evaluation of the different models and approaches is done with a 6-fold cross-validation using the task training set. Each fold consists of about 1,000 tweets as test while the others are used as train and validation (500 tweets). The performance score is obtained as the average of the six folds, so that the final evaluation is unbiased and independent as much as possible from the specific splits into train, validation and test.

In our setup we tested two models, first Multilingual BERT, covering 104 languages including Italian [2] and then AlBERTo, which was trained using the official BERT source code on 200M tweets in the Italian language. For the fine-tuning of AlBERTo we run it for 15 epochs, using a learning rate of 2e-5 with 1000 steps per loops on batches of 64 examples. Since AlBERTo performed better than multilingual BERT on each fold, it was included in the final system configuration for the task. The cross-validation over 6 folds using only the task training set with AlBERTo resulted in an average Macro-F1 of 83.12 for Run1 and 82.15 for Run2.

5.2 Data Preprocessing

The data, both from the dataset provided by the organisers and the silver one, are preprocessed as follows. First we split hashtags by adapting to Italian the Ekphrasis tool (Gimpel et al., 2010), which recognises the tokens in a hashtag based on Google n-grams. With the same tool we also normalise the text to replace all mentions to users and urls with <user> and <url> respectively. We also replace with a dedicated tag all the instances of *"money"*, *"time"*, *"date"* and in general any *"number"*. The emojis are replaced with their descriptions[3] in order to have a textual representation to be used with AlBERTo.

[2] with 12-layer, 768-hidden, 12-heads, 110M parameters

[3] manually translated to Italian from the English description at `rhttps://unicode.org/emoji/charts/full-emoji-list.html`.

		Hate class			Non-hate class			Macro Avg.
DocType.	System	Precision	Recall	F1	Precision	Recall	F1	F1
Tweets	Run1	0.7237	0.7958	0.758	0.7806	0.7051	0.7409	0.7495
	Run2	0.727	0.8006	0.762	0.7855	0.7083	0.7448	**0.7534**
	baselineMF	0	0	0	0.5075	1.000	0.6733	0.3366
	baselineSVM	0.7096	0.7347	0.7219	0.7334	0.7082	0.7206	0.7212
	best system							*0.8088*
News	Run1	0.6833	0.453	0.5448	0.7395	0.8808	0.804	0.6744
	Run2	0.6911	0.5193	0.593	0.7609	0.8683	0.8111	**0.702**
	baselineMF	0	0	0	0.638	1.000	0.7789	0.3894
	baselineSVM	0.6071	0.3756	0.4641	0.7087	0.862	0.7779	0.621
	best system							*0.7744*

Table 1: Results of the two submitted runs for Task A on tweets and on news headlines. BaselineMF = most-frequent baseline; baselineSVM = linear SVM with unigrams, char-grams and TF-IDF representation

6 Evaluation

We submitted two runs each for the in-domain (tweets) and out-of-domain (news headlines) text types in Task A. The results obtained on the test set are reported in Table 1 and compared with two baselines provided by the task organisers, one obtained by always assigning the most frequent label (i.e. non-hateful), and the other by training an SVM classifier with unigrams, char-grams and TF-IDF representation as features. We also compare our results with the top-ranked system in each subtask (additional details on such systems have not been disclosed at the moment of writing).

As expected, on out-of-domain data (news headlines) we obtain lower results than on tweets, since the training set is retrieved exclusively from Twitter. Furthermore, our approach does not include any specific tuning aimed at treating news headlines differently from tweets. On the contrary, the additional data used for self-training are all gathered from Twitter, which may negatively affect performance on out-of-domain data.

On both document types, Run2 performs better than Run1, showing that our oversampling strategy to reduce the weight of silver data is effective. However, results obtained with 6-fold cross-validation only on the training set were significantly higher, both with macro F1 > 0.80. This may be explained by the fact that, as pointed out by the task organisers, tweets from the test set were collected in a different time period than those in the training set. This will likely make the two sets different in terms of topics.

Run 1		Actual Values	
		non-hate	hate
Predicted	non-hate	452	127
	hate	189	495
Run 2		Actual Values	
		non-hate	hate
Predicted	non-hate	454	124
	hate	187	498

Table 2: Confusion matrix on *tweets* results

We report in Table 2 and 3 the confusion matrix showing the number of true positives and negatives, and false positives and negatives obtained with the two runs on tweets and news headlines. While on tweets the performance on the hate class is overall better, in particular concerning recall, this does not apply to news headlines, with a low recall for the hate class. The reason for this low score lies in the different linguistic expressions connected with hate between tweets and headlines: while in tweets they are more direct, and more frequently connected with profanities that a classifier can easily recognise, hateful content in news headlines is usually expressed in more subtle ways. As an example, we report below two headlines misclassified by our system. The first one (i) was classified as non-hateful, even if it conveys hateful content. The second one (ii) was instead classified as hateful, although it is not:

Run 1		Actual Values	
		non-hate	hate
Predicted	non-hate	281	99
	hate	38	82
Run 2		Actual Values	
		non-hate	hate
Predicted	non-hate	277	87
	hate	42	94

Table 3: Confusion matrix on *news headlines* results

i) *Sea Watch, l'ultima presa in giro degli immigrati all'Italia: i minori nati tutti lo stesso giorno* (EN: Sea Watch, migrants making fun of Italy: all underage migrants born on the same day)

ii) *Matera, Salvini contestato durante il comizio. E lui risponde: "Bravi, avete vinto dieci immigrati da mantenere"* (EN: Matera, Salvini challenged at a rally, and he replies: "Congratulations, you won ten migrants to pay for")

Both examples have a similar structure, are written in standard Italian and mention migrants. Furthermore, the second example reports a hateful direct speech, but since it is only reported it does not mean that the journalist agrees with what was said by the politician Matteo Salvini.

7 Conclusions

In this paper we described the system developed by the DH-FBK team to participate in the HaSpeeDe shared Task A. We submitted two runs, both based on AlBERTo and using in-domain silver data as additional training data in a self-learning framework. The only difference between the two configurations is that, for Run2, the task training data were repeated five times, to balance the weight of silver data.

Our evaluation shows that, both in a cross-validation setting and on the task test set, over-sampling has a positive effect on the classification results. As expected, performance on in-domain data (i.e. training and testing on tweets) is better than on out-of-domain data (i.e. training on tweets and testing on news headlines). In the future, we may try to address this issue by including as silver data also news headlines, so that also the specificity of this kind of text is taken into account. For a better data quality, it may be useful to select only the silver instances that have been automatically classified with high confidence.

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 the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task.

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.

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 Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

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.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task learning in deep neural networks at EVALITA 2018. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

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 the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. Cross-platform evaluation for italian hate speech detection. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology*, 20(2):10:1–10:22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, June.

Jérôme Ferret, Mario Laurent, Daniela Andreatta, Andrea Di Nicola, Elisa Martini, M Guerini, S Tonelli, Georgios Antonopoulos, and Parisa Diba. 2019. Hatemeter d18: Training module a for academics and research organisations.

Paula Fortuna, Ilaria Bonavita, and Sérgio Nunes. 2018. Merging datasets for hate speech classification in italian. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2010. Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon University, School of Computer Science.

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 the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

Marco Polignano and Pierpaolo Basile. 2018. Hansel: Italian hate speech detection through ensemble learning and deep neural networks. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019a. Hate speech detection through alberto italian language understanding model. In Mehwish Alam, Valerio Basile, Felice Dell'Orletta, Malvina Nissim, and Nicole Novielli, editors, *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AIIA 2019), Rende, Italy, November 19th-22nd, 2019*, volume 2521 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019b. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

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 the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

Dirk von Grunigen, Ralf Grubenmann, Fernando Benites, Pius Von Daniken, 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)*.

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)*, pages 1 – 10, Vienna, Austria. Austrian Academy of Sciences.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.