Hate Speech Detection through AlBERTo Italian Language Understanding Model

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro

University of Bari, Dept. Computer Science, E.Orabona 4, 70125, Bari, Italy
marco.polignano@uniba.it, pierpaolo.basile@uniba.it,
marco.degemmis@uniba.it, giovanni.semeraro@uniba.it

Abstract. The task of identifying hate speech in social networks has recently attracted considerable interest in the community of natural language processing. This challenge has great importance for identifying cyberattacks on minors, bullying activities, misogyny, or other kinds of hate discriminations that can cause diseases. Identifying them quickly and accurately can, therefore, help to solve situations that are dangerous for the health of the attacked people. Numerous national and international initiatives have addressed this problem by providing many resources and solutions to the problem. In particular, we focus on the Hate Speech Detection evaluation campaign (HaSpeeDe) held at Evalita 2018. It proposes an evaluation campaign with the aim of developing strategies for identifying hate speeches on Twitter and Facebook written in the Italian language. The dataset released for the task has been used by the classification approach proposed in this work for demonstrating that it is possible to solve the task efficiently and accurately. Our solution is based on an Italian Language Understanding model trained with a BERT architecture and 200M of Italian Tweets (AlBERTo). We used AlBERTo for fine-tuning a classification model of hate speech, obtaining state of the art results considering the best systems presented at the HaSpeeDe workshop. In this regard, AlBERTo is here proposed as one of the most versatile resources to be used for the task of classification of Social Media Textual contents in the Italian Language. The claim is supported by the similar results obtained by AlBERTo in the task of sentiment analysis, and irony detection demonstrated in previous works. The resources need for fine-tuning AlBERTo in these classification tasks are available at: https://github.com/marcopoli/AlBERTo-it

Keywords: Language Understanding Model · AlBERTo · Hate Speech · Classification · Machine Learning · Deep Learning

1 Introduction and Motivations

Hate speeches are characterized by their wide diffusion on the web and by the anonymity of the author, which makes this type of problem risky and relevant for the community. These messages can be against groups of people, such as those concerning discriminations about religion, race, and disability, or to a specific

person. In addition, hate messages are characterized by different facets that are very different from each other and which give rise to a wide and varied problem. The interpretation of a message as hate or not is subject to a strong cultural and social influence by making the same message hatefully for some subjects (e.g. them from a specific country) and not hatefully for others. Hate Speech (HS) is, consequently, a multi-faceted problem with strong cultural and social intersections. The lexicon used in that messages is difficult to be found in a standard dictionary and with many lexical variations making approaches of classification based only on dictionaries unsuccessful. Therefore, the automatic identification of hate messages is often a complex and intrinsically multidisciplinary task, including the research areas of natural language processing (NLP), psychology, law, social sciences, and many more. The hate speech detection is a challenging task that gains hight interest by private industries and public institutions to be able to remove potentially illegal contents quickly from the Web and to reduce the connected risk to remove legal content unjustly. This has made it interesting for us to apply an innovative classification model based on a language understanding model for the Italian language (AlBERTo [29].) to obtain promising results for the task.

The classification model of this work is part of a wider national project, "Contro l'odio"[1], that aims to monitor, classify and summarize in statistics the hate messages in Italian identified via Twitter. "Contro l'odio" is a project for countering and preventing racist discrimination and HS in Italy, in particular focused against immigrants. On the one hand, the project follows and extends the research outcomes emerged from the 'Italian Hate Map project' [23], whose goal was to identify the most-at-risk areas of the Italian country, that is to say, the areas where the users more frequently publish hate speech, by exploiting semantic analysis and opinion mining techniques. On the other hand, "Contro l'odio" benefits from the availability of annotated corpora for sentiment analysis, hate speech detection and related phenomena such as aggressiveness and offensiveness, to be used for training and tuning the HS detection tools [31,27]. The project brings together the competences and active participation of civil society organizations Acmos[2] and Vox[3], and two academic research groups, respectively from the University of Bari and Turin.

2 Related Work

The interest of the scientific community in the task of identifying hate speech and related phenomena such as misogyny, cyberbullying, and abusive language has been growing since 2016. Events such as HatEval 2019 [3], AMI at IberEval 2018 [15], HaSpeeDe 2018 [9] and AIM 2018 [16] at EVALITA 2018 have contributed to the emergence of a strong community of reference, methods, resources and tools to address this complex task. For what concerns Italian a few resources

[1] `https://controlodio.it/`
[2] `http://acmos.net/`
[3] `http://www.voxdiritti.it/`

have been recently developed drawn from Twitter [31,27] and Facebook [13], where the annotation of hateful contents also extends the simple markup of HS. A multilingual lexicon of hate words has also been developed [5] called Hurtlex[4]. It is divided into 17 categories such as homophobic slurs, ethnic slurs, genitalia, cognitive and physical disabilities, animals, and more.

A recent survey of state of the art approaches for hate speech detection is provided by Schmidt et al. [33]. The most common systems of speech detection are based on algorithms of text classification that use a representation of contents based on "surface features" such as them available in a bag of words (BOW) [11,37,36,34]. A solution based on BOW is efficient and accurate, especially when n-grams have been extended with semantic aspects derived by the analysis of the text. [11] describe an increase of the classification performances when features such as the number of URLs, punctuations and not English words are added to the vectorial representation of the sentence. [35] proposed, instead, to add as a feature the number of positive, negative, and neutral words found in the sentence. This idea demonstrated that the polarity of sentences positively supports the classification task. These approaches suffer from the lack of generalization of words contained into the bag of words, especially when it is created through a limited training set. In particular, terms found in the test sentences are often missing in the bag. More recent works have proposed word embeddings [19] as a possible distributional representation able to overcome this problem. This representation has the advantage to transform semantically similar words into a similar numerical vector (e.g. Word2Vec). Word embeddings are consequently used by classification strategies such as Support Vector Machine and recently by deep learning approaches such as deep recurrent neural networks [20].

Limits of such technologies as Word2Vec [22], Glove [25], and FastText [8] fall into the lack of use of context of terms when such representation is built (context-free). This means that each term has only a single wordembedding representation in the distribution space, and different concepts related to the same term are not represented. New strategies such as ELMO [26], GPT/GPT-2 [30], and BERT [14] overcome this limit by learning a language understanding model for a contextual and task-independent representation of terms. In particular, these models are trained to predict the totality or a span of the starting sentence. This allows obtaining a model able to predict, from a specific context (often both previous and subsequent), the most probable word from its vocabulary. Recently, several articles have demonstrated the effectiveness of this technique in almost all NLP tasks in the English language, and recently, some multilingual models have been distributed. This entails significant limitations related to the type of language learned (related to the document style) and the limit of vocabulary extracted. These reasons have led us to create the equivalent of the BERT model for the Italian language and specifically on the language style used on social networks: **alBERTo** [29].

[4] `http://hatespeech.di.unito.it/resources.html`

The classifier proposed in this work about HS is based on AlBERTo, demonstrating that its fine-tuned version is suitable for the task and it obtains better results than them presented ad HaSpeeDe 2018 evaluation campaign.

3 AlBERTo-HS classification model

The aim of this work is to create a classification model able to accurately classify HS contents written in the Italian Language on Social Network such as Facebook and Twitter. The analysis of state of the art shown that the main strategies for facing these challenges, on the English language, are currently based on a pre-trained language understanding models. Them, even in their multilingual version, are not suitable for an use with data completely in a single language and with a writing style different from that of books and encyclopedic descriptions. It is well known that the language used on social networks is different from the formal one as consequence of the presence of mentions, uncommon terms, links, and hashtags that are not present elsewhere. AlBERTo [29] wants to be the first Italian language understanding model to represent a style of writing of social networks, Twitter in particular, written in Italian. The fine-tuned classification model proposed in this work is based AlBERTo derived by the software code distributed through GitHub by Devlin et al. [14] [5] under the concession of Google. It has been suitably modified to be learned without consequences on text spans containing typical social media characters including emojis.

The core deep learning structure of BERT and AlBERTo is a $12x$ Transformer Encoder, where for each input, a percentage of terms is hidden and then predicted for optimizing network weights in back-propagation. This strategy of learning is commonly named "masked learning". In AlBERTo we implement only the "masked learning" strategy, excluding the one based on "next following sentence". This is a crucial aspect to be aware of because, in the case of tweets, we do not have cognition of a flow of tweets as it happens in a dialog. For this reason, we are sure enough that our AlBERTo is not suitable for the task of question answering where this property is essential to have been learned by the model. On the contrary, it is good enough to be used in tasks of classification and predictions. In order to tailor the tweet text to BERT's input structure, it has been necessary to carry out pre-processing operations. More specifically, using Python as the programming language, two libraries were mainly adopted: Ekphrasis [6] and SentencePiece [6] [18]. Ekphrasis is a famous tool for performing an NLP pipeline on text extracted from Twitter. It has been used for:

- Normalizing URL, emails, mentions, percents, money, time, date, phone numbers, numbers, emoticons;
- Tag and unpack hashtags.

The normalization phase consists in replacing the term with a fixed one in the style of $< [entity \quad type] >$. The tagging phase consists of annotating hashtags

[5] https://github.com/google-research/bert/
[6] https://github.com/google/sentencepiece

by two tags < *hashtag* > ... < /*hashtag* > representing its beginning and end in the sentence. Whenever possible, the hashtag has been unpacked into known words. For making the text clean and easily readable by the network, it has been returned to its lowercase form and all characters except emojis, !, ? and accented characters have been deleted.

SentencePiece is a segmentation algorithm used for learning in an unsupervised and language independent way the best strategy for splitting text into terms for language models. It can process till 50k sentences per seconds and to generate an extensive vocabulary. It includes in it the most common terms in the training set and the subwords which occur in the middle of words, annotating them with '##' in order to be able to encode also slang, incomplete or uncommon words. SentencePiece produced also a tokenizer used for generating a list of tokens for each tweet lately processed by the BERT *"create_pretraining_data.py"* module. The dataset used for the learning phase of AlBERTo is TWITA [4] a huge corpus of Tweets in the Italian language collected from February 2012 to September 2015 from Twitter official streaming API. In our configuration, we randomly selected 200.000.000 of Tweets removing re-tweets , and processing them with the pipeline of pre-processing previously described. The AlBERTo classification model is the basis for any single-label or multi-label classification task. For the specific task of content classification of Hate speech, we will carry out a subsequent phase of fine-tuning and adaptation of the model to domain-specific data. This allows us to obtain a classifier that exploits the language knowledge obtained during the learning phase on the generic data and the specific domain characteristics learned during the fine-tuning phase. The fine-tuning phase is configured as a new training of AlBERTo with a number of epochs sufficiently small not to overfit the model on the new data provided (usually from 3 to 15 epochs). This process allows us to vary the weights of the last layers of the model in order to predict correctly the content provided in the testing phase. We named the fine-tuned version of AlBERTo for Hate Speech as **AlBERTo-HS**.

4 Evaluation

In order to evaluate Alberto-HS with contents produced by real users on social networks, written in the Italian language, we decided to use the data released for the evaluation campaign HaSpeeDe [9] at EVALITA 2018. This choice was made considering that most of the available state of the art datasets are in English or focused only on data collected from a single social media site such as Facebook, Twitter, and others. The HaSpeeDe evaluation campaign was carried out by dividing the problem into four different tasks:

- **HaSpeeDe-FB**: where the goal is to train the model and predict if the contents are HS on data extracted from Facebook;
- **HaSpeeDe-TW**: where the goal is to train the model and predict if the contents are of HS on data extracted from Twitter;

- **Cross-HaSpeeDe_FB**: where the goal is to train the model on data collected from Facebook and predict if the contents are of HS on data extracted from Twitter;
- **Cross-HaSpeeDe_TW**: where the goal is to train the model on data collected from Twitter and predict if the contents are of HS on data extracted from Facebook;

It is interesting to note that in the first two tasks, the model must be able to classify data coming from the same information source as the training phase. Unlike the two "Cross" tasks, the data to be classified are different from those used for the test, making the task of the classifier more challenging due to the differences in writing styles of the two platforms. In fact, not only are twitter data shorter, containing mentions, hashtags, and retweets, but overall, they are also less HS than Facebook data (only 32% compared to 68% for Facebook).

4.1 Dataset and Metrics

Facebook dataset is collected from public pages on Facebook about newspapers, public figures, artists and groups on heterogeneous topics. More than 17,000 comments were collected from 99 posts and subsequently annotated by 5 bachelor students. The final dataset released consists of 3000 training phrases (1618 not HS, 1382 HS) and 1000 test phrases (323 not HS, 677 HS).

Twitter dataset is part of the Hate Speech Monitoring program, coordinated by the Computer Science Department of the University of Turin with the aim at detecting, analyzing and countering HS with an inter-disciplinary approach [10]. Data were collected using keywords related to the concepts of immigrants, Muslims and Rome. Data are annotated partly by experts and partly by Figure Eight contributors. Also for this dataset 3000 training tweets were released (2028 not HS and 972 HS) and 1000 test tweets (676 not HS and 324 HS).

The evaluation metrics used in HaSpeeDe campaign are the Precision, Recall and F1-measure classics. Since the two classes (HS and not HS) are unbalanced within the datasets, the F1 metric has been calculated separately on the two classes and then macro-averaged.

4.2 AlBERTo-HS fine-tuning

We fine-tuned AlBERTo two different times, in order to obtain one classifier for each different dataset available as a training set. In particular, we created one classifier for the HaSpeeDe-FB and the Cross-HaSpeeDe_FB tasks using Facebook training data and one for the HaSpeeDe-TW and the Cross-HaSpeeDe_TW using the Twitter training set. The fine-tuning learning phase has been run for 15 epochs, using a learning rate of 2e-5 with 1000 steps per loops on batches of 512 examples. The fine-tuning process was last $\sim$ 4 minutes every time.

4.3 Systems and baseline

HaSpeeDe has received strong participation from the scientific community and therefore a large number of solutions to the task have been proposed [9].

GRCP [24] The authors developed a Bi-LSTM Deep Neural Network with an Attention-based mechanism that allows to estimate the importance of each word; the weight vector is then used with another LSTM model to classify the text.

HanSEL [28] The system proposed is based on an ensemble of three classification strategies (Support Vector Machine with RBF kernel, Random Forest and Deep Multilayer Perceptron), mediated by a majority vote algorithm. The social media text is represented as a concatenation of word2vec vectors and a TF-IDF bag of words.

InriaFBK [21] The authors implemented three different classifier models: RNN, n-gram based and linear SVC.

ItaliaNLP [12] Participants used a newly-introduced model based on a 2-layer BiLSTM which exploits multi-task learning with additional data from the 2016 SENTIPOLC task [2].

Perugia [32] The participants' system uses a document classifier based on a SVM algorithm. The features used by the system are a combination of FastText word embeddings and other 20 syntactical features extracted from the text.

RuG [1] The authors proposed two different classifiers: a SVM based on linear kernel and an ensemble system composed of an SVM and a CNN combined by a logistic regression meta-classifier.

sbMMMP The authors tested two different systems. The first one is based on an ensemble of CNNs, whose outputs are then used as features by a meta-classifier for the final prediction. The second system uses a combination of CNN and a GRU.

StopPropagHate [17] The authors use a classifier based on RNN 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 [7] 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.

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

4.4 Discussion of results

The evaluation of the results obtained by the AlBERTo-HS classifier proposed in this work was carried out using the official evaluation script released at the end of the campaign [7]. Consequently, all the results obtained are replicable and comparable with those present in the final ranking of HaSpeeDe.

[7] http://www.di.unito.it/ tutreeb/haspeede-evalita18/data.html

Table 1. Results of the HaSpeeDe-FB task

	NOT HS			**HS**			
	Precision	Recall	F-score	Precision	Recall	F-score	***Macro-Avg F-score***
most_freq							0.2441
AlBERTo-HS	**0.8603**	0.7058	**0.7755**	0.8707	**0.9453**	**0.9065**	**0.8410**
ItaliaNLP 2	0.8111	0.7182	0.7619	0.8725	0.9202	0.8957	0.8288
InriaFBK 1	0.7628	0.6873	0.7231	0.8575	0.898	0.8773	0.8002
Perugia 2	0.7245	0.6842	0.7038	0.8532	0.8759	0.8644	0.7841
RuG 1	0.699	0.6904	0.6947	0.8531	0.8581	0.8556	0.7751
HanSEL	0.6981	0.6873	0.6926	0.8519	0.8581	0.855	0.7738
VulpeculaTeam	0.6279	0.7523	0.6845	0.8694	0.7872	0.8263	0.7554
RuG 2	0.6829	0.6068	0.6426	0.8218	0.8655	0.8431	0.7428
GRCP 2	0.6758	0.5294	0.5937	0.7965	0.8788	0.8356	0.7147
StopPropagHate 2	0.4923	0.6965	0.5769	0.8195	0.6573	0.7295	0.6532
Perugia 1	0.3209	0.9907	0.4848	0	0	0	0.2424

Table 2. Results of the HaSpeeDe-TW task

	NOT HS			**HS**			
	Precision	Recall	F-score	Precision	Recall	F-score	**Macro-Avg F-score**
most_freq							0.4033
AlBERTo-HS	0.8746	0.8668	0.8707	0.7272	0.7407	**0.7339**	**0.8023**
ItaliaNLP 2	0.8772	0.8565	0.8667	0.7147	0.75	0.7319	0.7993
RuG 1	0.8577	0.8831	0.8702	0.7401	0.6944	0.7165	0.7934
InriaFBK 2	0.8421	0.8994	0.8698	0.7553	0.6481	0.6976	0.7837
sbMMMP	0.8609	0.852	0.8565	0.6978	0.7129	0.7053	0.7809
VulpeculaTeam	0.8461	0.8786	0.8621	0.7248	0.6666	0.6945	0.7783
Perugia 2	0.8452	0.8727	0.8588	0.7152	0.6666	0.69	0.7744
StopPropagHate 2	0.8628	0.7721	0.8149	0.6101	0.7438	0.6703	0.7426
GRCP 1	0.7639	0.8713	0.814	0.62	0.4382	0.5135	0.6638
HanSEL	0.7541	0.8801	0.8122	0.6161	0.4012	0.4859	0.6491

Table 3. Results of the Cross-HaSpeeDe_FB task

	NOT HS			HS			
	Precision	Recall	F-score	Precision	Recall	F-score	**Macro-Avg F-score**
most_freq							0.4033
InriaFBK 2	0.8183	0.6597	0.7305	0.4945	0.6944	0.5776	0.6541
VulpeculaTeam	0.8181	0.639	0.7176	0.483	0.7037	0.5728	0.6452
Perugia 2	0.8503	0.5547	0.6714	0.4615	0.7962	0.5843	0.6279
ItaliaNLP 1	0.9101	0.4644	0.615	0.4473	0.9043	0.5985	0.6068
GRCP 2	0.7015	0.7928	0.7444	0.4067	0.2962	0.3428	0.5436
RuG 1	0.8318	0.4023	0.5423	0.3997	0.8302	0.5396	0.5409
AlBERTo-HS	0.8955	0.2662	0.4104	0.3792	0.9351	0.5396	0.4750
HanSEL	0.7835	0.2677	0.3991	0.3563	0.8456	0.5013	0.4502
StopPropagHate	0.6579	0.3727	0.4759	0.3128	0.5956	0.4102	0.443

Table 4. Results of the Cross-HaSpeeDe_TW task

	NOT HS			**HS**			
	Precision	Recall	F1-score	Precision	Recall	F1-score	**Macro F1-score**
most_freq							0.2441
ItaliaNLP 2	0.5393	0.7647	0.6325	0.8597	0.6883	0.7645	0.6985
AlBERTo-HS	0.5307	0.7492	0.6213	0.8511	0.6838	0.7583	0.6898
InriaFBK 2	0.5368	0.6532	0.5893	0.8154	0.7311	0.771	0.6802
VulpeculaTeam	0.453	0.7461	0.5637	0.8247	0.5701	0.6742	0.6189
RuG 1	0.4375	0.6934	0.5365	0.7971	0.5745	0.6678	0.6021
HanSEL	0.3674	0.8235	0.5081	0.7934	0.3234	0.4596	0.4838
Perugia 2	0.3716	0.9318	0.5313	0.8842	0.2481	0.3875	0.4594
GRCP 1	0.3551	0.8575	0.5022	0.7909	0.257	0.3879	0.4451
StopPropagHate	0.3606	0.9133	0.517	0.8461	0.2274	0.3585	0.4378

From the previous tables of results, it is possible to observe how AlBERTo-HS succeeds in obtaining a state of the art results for two tasks out of four. The differences with other systems proposed in the evaluation campaign are about its simplicity to be applied. A simple fine-tuning phase of AlBERTo on domain data allows us to obtain very encouraging results. It is therefore interesting to note that the entire process of pre-processing and fine-tuning lasts a few minutes, and it can be used for obtaining excellent results for a wide variety of classification tasks. In particular, the model is able to adapt in an excellent way to annotated data (with the risk of overfitting) producing excellent results if used in the same application domain of the tuning phase. This is the case with the results obtained for the HaSpeeDe-FB and HaSpeeDe-TW tasks.

Looking at the results obtained for the classification of data coming from Facebook (Tab. 1), it is possible to observe how the classifier is able to capture the characteristics of the social language through the fine-tuning phase. In particular, it is able to move its learned weights from them obtained parsing the original training language based on Twitter to the one used on Facebook. AlBERTo-HS obtains better performances than those of other participants in the evaluation campaign, with regard to the precision in identifying the posts not hate (0.8603), and the recall of those of hate (0.9453). The high value of recall for hate messages allows us to assume that, on Facebook, they are characterized by specific thematics that make the classification task more inclusive at the cost of accuracy, especially when not explicit hate messages are faced. As an example, the message "Comunque caro Matteo se non si prendono provvedimenti siamo rovinati." is classified as a hate message even if the annotators have considered it to be not a hate message. In this example, it is clear that a basis of hate is present in the ideas of the writer, even if it is not complicated by what he writes. In other cases, words like "severe" have tricked the model into classifying clearly neutral messages like the following as hate messages: "Matteo sei la nostra voce!!! Noi donne non possiamo fare un cavolo! !! Leggi più severe!". Nevertheless, the average F1 score higher than 0.8410, show us that, unlike in Twitter, the use of more characters available for writing allows people to be more verbose and, therefore, more comfortable to identify. Table 2 shows the results obtained for the classification of tweets. Here the values are not so different from the first in the ranking during the evaluation campaign Haspeede even if the average value of F1 obtained of 0.8023 proves to be the best. This suggests that the presence in the tweets of particular characters and implicitly of hate, the brevity of the latter, and the increase in the number of ironic tweets make the task more complicated than the previous one.

As far as "Cross" classification problems are concerned, the results are not guaranteed. In Tab. 3 it can be observed that the model has not been able to correctly abstract from the domain data, obtaining not very good results for the classification in a different domain. In particular, the model trained on Facebook is able to obtain a score of 0.4750 of F1 on Twitter test data. A similar situation is repeated for the results in Tab. 4 where for the task Cross-HaSpeeDe_TW the model is able to generalize slightly better than before but still gets the second

place in the ranking. These results confirm the difficulty of the Cross tasks and the drop in performance that is obtained through a transfer-learning strategy like the one adopted here. The great differences in writing styles used on the two social networks do not allow the model to adapt properly to the domain of application if fine-tuned on different stylistic data. So that AlBERTo is not able to grasp those particularities of the language to be used in the classification phase.

In any case, we want to observe how it has been possible to obtain an excellent result of classification by merely carrying out a phase of fine-tuning on the model. To this end, we will consider as future works those of making a further comparison with other language understanding models such as GPT2, XLNet, RoBERTa trained on the Italian language with the aim of verifying if they can be more robust to the changes in the writing style of the text to be classified.

5 Conclusion

The problem of hate speech is strongly perceived in online communities because of its repercussions on the quality of life of hate victims. It is therefore of great interest to both public and private organisations to be able to quickly identify and remove hate messages. Numerous national and international initiatives have been carried out in recent years, especially for the English language, leaving the Italian language with few resources to address the problem. In this work we have proposed a simple model of classification obtainable through a quick fine-tuning phase of a wider language understanding model pre-trained on the Italian language (AlBERTo). This model was evaluated on the data released for the HaSpeeDe evaluation campaign held at the EVALITA 2018 workshop. Data containing phrases extracted from Facebbok and Twitter were classified according to four different tasks. The first two involved training the model on data from the same domain as the test data. On the contrary, the last two "Cross" tasks involved a classification on data from a domain different from the training one. The results obtained showed excellent performances when the model is evaluated on data coming from the same distribution of training data. On the contrary, good performances in this transfer learning task are not guaranteed due to the great stylistic differences of the language used on different online platforms such as Facebook and Twitter. Future work will focus on the possibility of learning a model that includes data from different online sources so as to make it more complete and robust to stylistic variations.

6 Acknowledgment

This work is funded by project "DECiSION" codice raggruppamento: BQS5153, under the Apulian INNONETWORK programme, Italy.

References

1. Bai, X., Merenda, F., Zaghi, C., Caselli, T., Nissim, M.: Rug@ evalita 2018: Hate speech detection in italian social media. In: EVALITA@ CLiC-it (2018)
2. Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V.: Overview of the evalita 2016 sentiment polarity classification task. In: Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016) (2016)
3. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M.: 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. pp. 54–63. Association of Computational Linguistics (2019)
4. Basile, V., Lai, M., Sanguinetti, M.: Long-term social media data collection at the university of turin. In: Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). pp. 1–6. CEUR-WS (2018)
5. Bassignana, E., Basile, V., Patti, V.: Hurtlex: A multilingual lexicon of words to hurt. In: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018. CEUR Workshop Proceedings, vol. 2253. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2253/paper49.pdf
6. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 747–754. Association for Computational Linguistics, Vancouver, Canada (August 2017)
7. Bianchini, G., Ferri, L., Giorni, T.: Text analysis for hate speech detection in italian messages on twitter and facebook. 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. (2018), http://ceur-ws.org/Vol-2263/paper043.pdf
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
9. Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., Maurizio, T.: Overview of the evalita 2018 hate speech detection task. In: Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018). vol. 2263, pp. 1–9. CEUR (2018)
10. Bosco, C., Viviana, P., Bogetti, M., Conoscenti, M., Ruffo, G., Schifanella, R., Stranisci, M.: 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 (2017)
11. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). pp. 71–80. IEEE (2012)

12. Cimino, A., De Mattei, L., Dell'Orletta, F.: Multi-task learning in deep neural networks at evalita 2018. In: EVALITA@ CLiC-it (2018)
13. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate Me, Hate Me Not: Hate Speech Detection on Facebook. In: Proceedings of the First Italian Conference on Cybersecurity (ITASEC17) (2017)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://www.aclweb.org/anthology/N19-1423
15. Fersini, E., Nozza, D., Rosso, P.: Overview of the evalita 2018 task on automatic misogyny identification (ami). In: Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR. org (2018)
16. Fersini, E., Nozza, D., Rosso, P.: Overview of the evalita 2018 task on automatic misogyny identification (AMI). In: Caselli, T., Novielli, N., Patti, V., Rosso, P. (eds.) 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. CEUR Workshop Proceedings, vol. 2263. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2263/paper009.pdf
17. Fortuna, P., Bonavita, I., Nunes, S.: 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. (2018), http://ceur-ws.org/Vol-2263/paper037.pdf
18. Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. arXiv preprint arXiv:1804.10959 (2018)
19. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning. pp. 1188–1196 (2014)
20. Mehdad, Y., Tetreault, J.: Do characters abuse more than words? In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 299–303 (2016)
21. Michele, C., Stefano, M., Pinar, A., Sprugnoli, R., Elena, C., Sara, T., Serena, V.: Comparing different supervised approaches to hate speech detection. In: EVALITA 2018. pp. 230–234. aAccademia University Press (2018)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
23. Musto, C., Semeraro, G., de Gemmis, M., Lops, P.: 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. pp. 307–308. ACM (2016)
24. De la Peña Sarracén, G.L., Pons, R.G., Muñiz-Cuza, C.E., Rosso, P.: Hate speech detection using attention-based lstm. In: EVALITA@ CLiC-it (2018)
25. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

26. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations pp. 2227–2237 (Jun 2018). https://doi.org/10.18653/v1/N18-1202, https://www.aclweb.org/anthology/N18-1202
27. Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., Bosco, C.: Hate speech annotation: Analysis of an italian twitter corpus. In: Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017. CEUR Workshop Proceedings, vol. 2006. CEUR-WS.org (2017), http://ceur-ws.org/Vol-2006/paper024.pdf
28. Polignano, M., Basile, P.: Hansel: Italian hate speech detection through ensemble learning and deep neural networks. In: EVALITA@ CLiC-it (2018)
29. Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., Basile, V.: AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019). CEUR (2019)
30. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
31. Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M.: An Italian Twitter Corpus of Hate Speech against Immigrants. In: Proceedings of the 11th Language Resources and Evaluation Conference 2018 (2018)
32. Santucci, V., Spina, S., Milani, A., Biondi, G., Di Bari, G.: Detecting hate speech for italian language in social media. In: EVALITA 2018, co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). vol. 2263 (2018)
33. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 1–10 (2017)
34. Sood, S., Antin, J., Churchill, E.: Profanity use in online communities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1481–1490. ACM (2012)
35. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., Hoste, V.: Detection and fine-grained classification of cyberbullying events. In: International Conference Recent Advances in Natural Language Processing (RANLP). pp. 672–680 (2015)
36. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media. pp. 19–26. Association for Computational Linguistics (2012)
37. Xu, J.M., Jun, K.S., Zhu, X., Bellmore, A.: Learning from bullying traces in social media. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. pp. 656–666. Association for Computational Linguistics (2012)