By1510 @ HaSpeeDe 2: Identification of Hate Speech for Italian Language in Social Media Data

Tao Deng, Yang Bai, Hongbing Dai†
School of Information Science and Engineering
Yunnan University, Yunnan, P.R. China
Dtao.top@gmail.com
baiyang.top@gmail.com
hbdai_it@126.com

Abstract

English. Hate speech detection has become a crucial mission in many fields. This paper introduces the system of team **By1510**. In this work, we participate in the HaSpeeDe 2 (Hate Speech Detection) shared task which is organized within E-valita 2020(The Final Workshop of the 7th evaluation campaign). In order to obtain more abundant semantic information, we combine the original output of BERT-Ita and the hidden state outputs of BERT-Ita. We take part in task A. Our model achieves an F1 score of 77.66% (6/27) in the tweets test set and our model achieves an F1 score of 66.38% (14/27) in the news headlines test set.

Italiano. *L' individuazione dell' incitamento allodio diventata una missione cruciale in molti campi. Questo articolo introduce il sistema del team By1510. In questo lavoro, partecipiamo al task HaSpeeDe 2 che stato organizzato allinterno di Evalita 2020. Per ottenere informazioni semantiche pi abbondanti abbiamo combinato loutput originale di BERT Ita e gli output di hidden state di BERT Ita. Il sistema presentato partecipa al task A. Il nostro modello ottiene un punteggio F1 di 77.66% (6/27) sui dati di test da Twitter e un punteggio F1 di 66.38% (14/27) sui dati di test contenenti titoli di quotidiano.*

1 Introduction

With the continuous development of computer and networks, social media users have increased year by year, social media has entered people's daily life and becomes an indispensable part. More and more people use the Internet to express their opinions and ideas on social media platforms. Some offensive, abusive, defamatory contents are easy to spread and incite hatred, and these negative contents can cause some bad effects. The simplest way is that people mark the report and then delete the system warning, which can not be solved efficiently. Therefore, an efficient way is urgently needed to eliminate these negative effects. This paper proposes a hate speech detection system, which can better detect and mark these annoying contents. The HaSpeeDe 2 (Sanguinetti et al., 2020) (Hate Speech Detection) shared task is organized within Evalita 2020 (Basile et al., 2020), the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian, which help to detect whether the Italian language on Twitter contains hate language, with the aim to reduce the spread of hate speeches and online harassment. (Waseem and Hovy, 2016)

In this paper, we take part in task A in the HaSpeeDe 2 task. The BERT model we use is dbmz[1] trained on Italian data. In order to obtain more abundant semantic information, we extract the state of hidden layer outputs and we provide a reference for the detection of the hate speech in the Italian language. The rest of the paper is organized as follows. Section 2 briefly shows the related work for the identification of hate speeches. Section 3 elaborates on our approach. It shows the data set officially provided and architecture of our model. Section 4 describes the hyper-parameters and our results. Finally, Section 5 concludes our work.

[1] https://huggingface.co/dbmdz

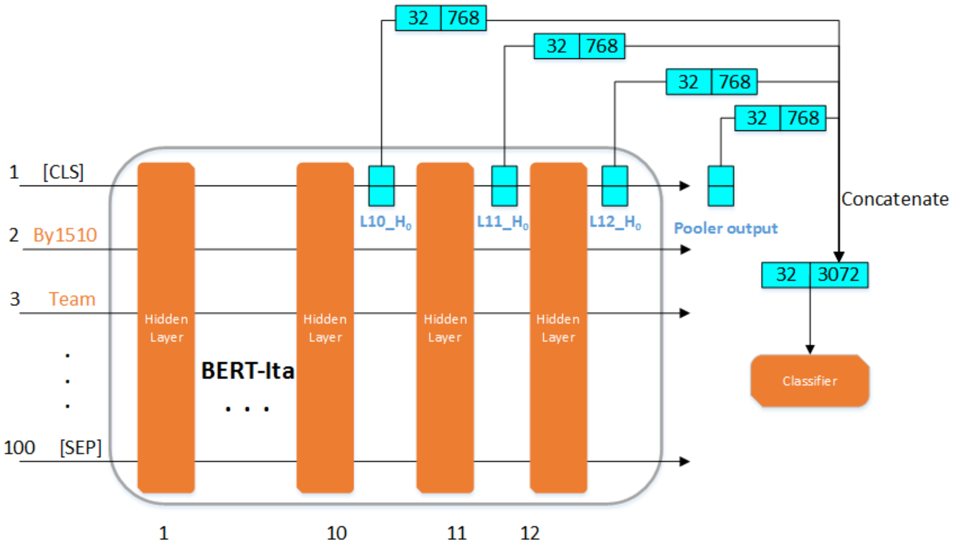

Figure 1: our model. $L12\_H_0$ is hidden-state of the first token of the sequence(CLS token) at the output of the 12th hidden layer of the BERT-Ita. Similarly, $L11\_H_O$ and $L10\_H_O$ are the 11th and 10th hidden layers outputs of BERT-Ita respectively. [32, 768]/[32, 3072] is the output shape (batch_size, hidden_size)

2 Related Work

Previously, machine learning (Davidson et al., 2017; MacAvaney et al., 2019a), Bayesian method (Miok et al., 2020; Fauzi and Yuniarti, 2018), support vector machine (MacAvaney et al., 2019b; Del Vigna12 et al., 2017), neural network (Badjatiya et al., 2017; Zhang et al., 2018) and other methods were proposed for the identification of hate speech. In the Hindi-English mixed language, (Bohra et al., 2018) et al. in parentheses used a supervised classification system to detect the hate speech in the text in the code-mixed language. The classification system used Character N-Grams, Word N-Grams, Punctuations, Negation Words, Lexicon and other feature vectors for classification and training. The accuracy could reach 71.7% with SVM, which proved to be a very effective method for classification tasks. In Danish language, (Sigurbergsson and Derczynski, 2019) developed four automatic classification systems to detect and classify hate speech in English and Danish, and proposed a method to automatically detect different types of the hate speech, which achieved good results for the detection of English and Danish hate speeches. In English language, (Aroyehun and Gelbukh, 2018) used a linear baseline classifier (nbsvm with n-grams) and improved deep neural network model.

For the Italian language, (Polignano et al., 2019) proposed an $AlBERTo$ model based on classifier integration, which was verified by cross validation on Facebook and Twitter data sets, and the effect was obvious in offensive words. (Corazza et al., 2018) used recurrent neural network, n-gram neural network and support vector machine to classify Twitter data sets, and its recurrent model had achieved good results. (Bianchini et al., 2018) proposed artificial neural network to annotate and classify 3000 message data from Facebook and Twitter, and achieved good results.

3 Methodology

3.1 Data Description

In this work, we take part in task A, which is a binary classification task aimed at determining the presence or the absence of hateful content in the text towards a given target (among Immigrants, Muslims or Roma people). The organizers provide the training set and test set. For the training set, it is from Twitter. For the test set, the organizers provide in-domain data and out-of-domain data, which come from Twitter and news headlines, respectively. It can be seen from Table 1 that the data set is slightly imbalanced.

3.2 Our approach

As the train data is very limited we resort to a transfer learning approach. That is, we take an NLP model pre-trained(Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019) on a large

	Hate Speech (HS)	No HS
train data	2766	4071
test data (tweets)	622	641
test data (news headlines)	181	319

Table 1: Distribution of data set in the Task A.

	Hyperparameters

Our Model	output_hidden_states=True
	max sequences length=100
	learning rate=1e-5
	adam_epsilon=1e-8
	per_gpu_train_batch_size=32
	gradient_accumulation_steps=1
	epoch=8
	dropout=0.1

Table 2: Hyperparameters of the model in our experiments.

corpus of texts and fine-tune it for a specific task at hand. In this work, we used BERT-base-Italian-uncased(BERT-Ita)[2] from Transformers library. It is trained on the recent Wikipedia dump and various texts from the OPUS corpora[3] collection. The final training corpus has a size of 13GB and 2050 million tokens. For classification tasks, the output of BERT-Ita ($pooler\ output$) is obtained by its last layer hidden state of the first token of the sequence (CLS token) further processed by a linear layer and a Tanh activation function. However, the pooler output is usually not a good summary of the semantic information. Therefore, we extract the hidden layer output of BERT-Ita to obtain more abundant semantic information.

(Jawahar et al., 2019) pointed that the hidden layer of BERT encodes a rich hierarchy of linguistic information, with surface features at the bottom layer, syntactic features in the middle layer and semantic features at the top layer. Therefore, we get abundant semantic information by extracting the extra semantic features which is the last three hidden layer outputs($L12\_H_0$, $L11\_H_0$ and $L10\_H_0$) of BERT-Ita. We propose the following model which is shown in Figure 1. In the model, we get $L12\_H_0$, $L11\_H_0$, $L10\_H_0$ from the top hidden layer of BERT-Ita. We concatenate $pooler\ output$, $L12\_H_0$, $L11\_H_0$ and $L10\_H_0$ into the classifier.

4 Experiments and Results

4.1 Preprocessing and Experiments Setup

In the experiment, we try to preprocess the text but we did not achieve the desired results. We find that after preprocessing the Twitter data, the F1-score of the model decreased on the validation set. We do not preprocess the data and we do not use an extra data set. In this work, the training set is split into the new training set and the validation set by using the Stratified 5-Fold Cross-validation[4].The random seed is set 42 in Cross-validation. Due to the imbalance of datasets, the Stratified 5-Fold Cross-validation ensures that the proportion of samples in each category in each fold data set remains unchanged. During the training, the best weight of the model is saved in 8 epochs. Table 2 shows the hyperparameters used in our model.

4.2 Results and analysis

In the experiment, we find that with the increase of the extra semantic features, the model can obtain more abundant semantic information. Table 3 shows the performance of the model for different semantic features after getting the labels of the test set.[5].

	Task A test set of tweets(100%)	
	No HS	HS
No HS	489	152
HS	119	503
	Task A test set of news headlines(100%)	
	No HS	HS
No HS	312	7
Hs	133	48

Table 4: The confusion matrix of BERT-Ita+L12_H_O in test sets.

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

[3] http://opus.nlpl.eu/

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html#sklearn.model_selection.StratifiedKFold

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

	Task A test set of tweets(100%)	Task A test set of news headlines(100%)
	Precision/Recall/Macro F1-score	Precision/Recall/Macro F1-score
BERT-Ita+L12_H_O	78.61/78.58/**78.54**	78.69/62.16/61.18
BERT-Ita+L12_H_O+L11_H_O	75.50/77.27/77.16	78.13/62.23/62.76
BERT-Ita+L12_H_O+L11_H_O+L10_H_O (Our submitted model)	77.80/77.72/77.66	72.07/65.74/**66.38**

Table 3: The performance of the model for these test sets.

	Task A test set of tweets(100%)	
	No HS	HS
No HS	478	163
HS	119	503
	Task A test set of news headlines(100%)	
	No HS	HS
No HS	289	30
Hs	107	74

Table 6: The confusion matrix of BERT-Ita+L12_H_O+L11_H_OF+L10_H_O in test sets.

	Task A test set of tweets(100%)	
	No HS	HS
No HS	463	178
HS	110	512
	Task A test set of news headlines(100%)	
	No HS	HS
No HS	310	9
Hs	128	53

Table 5: The confusion matrix of BERT-Ita+L12_H_O+L11_H_O in test sets.

The confusion matrices (actual values are represented by rows) are shown in Table 4, Table 5, Table 6. These tables show the performance of the model on the test set as the extra semantic features increase. In the tweets test set, we can see from these tables that the ability of the model to detect the hate speech is increasing as the extra semantic features increase. Similarly, in the news headlines test set, the ability of the model to detect the hate speech is also increasing. We think that with the increase of these extra semantic features, the model can learn more semantic information. In addition, we find that our model achieve good results on the tweets test set, but the results of our model are not good on the news headline data set. There are many differences between the syntactic features of tweets and news headlines. For example, there are many irregular expressions in tweets, while news expressions are very standard. Our model is only fine-tuned on the tweets data set, so we think this affects the performance of the model on other types of data.

5 Conclusion

In this work, this paper introduces the system proposed for HaSpeeDe 2 shared task for identifying and classifying hate speeches on social media. We enriched BERT-Ita with semantic information by extracting the extra semantic features. We find that with the increase of semantic information, the performance of the model for identifying the hate speech is also increasing. Finally, in the official evaluation, our model rank 6th (6/27) in the tweets test set and 14th (14/27) in the news headlines test set. In the future, we will focus on how to make the model learns more semantic information.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

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

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 Eval-*

uation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (E-VALITA 2020), Online. CEUR.org.

Giulio Bianchini, Lore nzo Ferri, and Tommaso Giorni. 2018. Text analysis for hate speech detection in i-talian messages on twitter and facebook. *EVALI-TA Evaluation of NLP and Speech Tools for Italian*, 12:250.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of peoples opinions, personality, and emotions in social media*, pages 36–41.

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing different supervised approaches to hate speech detection.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Fabio Del Vigna12, Andrea Cimino23, Felice DellOrletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

M Ali Fauzi and Anny Yuniarti. 2018. Ensemble method for indonesian twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, 11(1):294–299.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019a. Hate speech detection: Challenges and solutions. *P-loS one*, 14(8):e0221152.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019b. Hate speech detection: Challenges and solutions. *P-loS one*, 14(8):e0221152.

Kristian Miok, Blaz Skrlj, Daniela Zaharie, and Marko Robnik-Sikonja. 2020. To ban or not to ban: Bayesian attention networks for reliable hate speech detection. *arXiv preprint arXiv:2007.05304*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. Hate speech detection through alberto italian language understanding model. In *NL4AI@ AI* IA*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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 E-valuation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (E-VALITA 2020)*, Online. CEUR.org.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2019. Offensive language and hate speech detection for danish. *arXiv preprint arXiv:1908.04531*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.