Cross-Platform Evaluation for Italian Hate Speech Detection

Michele Corazza[†], **Stefano Menini**[‡],
Elena Cabrio[†], **Sara Tonelli**[‡], **Serena Villata**[†]
[†]Université Côte d'Azur, CNRS, Inria, I3S, France
[‡]Fondazione Bruno Kessler, Trento, Italy
michele.corazza@inria.fr
{menini,satonelli}@fbk.eu
{elena.cabrio,serena.villata}@unice.fr

Abstract

English. Despite the number of approaches recently proposed in NLP for detecting abusive language on social networks, the issue of developing hate speech detection systems that are robust across different platforms is still an unsolved problem. In this paper we perform a comparative evaluation on datasets for hate speech detection in Italian, extracted from four different social media platforms, i.e. Facebook, Twitter, Instagram and WhatsApp. We show that combining such platform-dependent datasets to take advantage of training data developed for other platforms is beneficial, although their impact varies depending on the social network under consideration.[1]

Italiano. *Nonostante si osservi un crescente interesse per approcci che identifichino il linguaggio offensivo sui social network attraverso l'NLP, la necessità di sviluppare sistemi che mantengano una buona performance anche su piattaforme diverse è ancora un tema di ricerca aperto. In questo contributo presentiamo una valutazione comparativa su dataset per l'identificazione di linguaggio d'odio provenienti da quattro diverse piattaforme: Facebook, Twitter, Instagram and WhatsApp. Lo studio dimostra che, combinando dataset diversi per aumentare i dati di training, migliora le performance di classificazione, anche se l'impatto varia a seconda della piattaforma considerata.*

1 Introduction

Given the well-acknowledged rise in the presence of toxic and abusive speech on social media platforms like Twitter and Facebook, there have been several efforts within the Natural Language Processing community to deal with such problem, since the computational analysis of language can be used to quickly identify offenses and ease the removal of abusive messages. Several workshops (Waseem et al., 2017; Fišer et al., 2018) and evaluation campaigns (Fersini et al., 2018; Bosco et al., 2018; Wiegand et al., 2018) have been recently organized to discuss existing approaches to hate speech detection, propose shared tasks and foster the development of benchmarks for system evaluation.

However, most of the available datasets and approaches for hate speech detection proposed so far concern the English language, and even more frequently they target a single social media platform (mainly Twitter). In low-resource scenarios it is therefore common to have smaller datasets for specific platforms, raising research questions such as: would it be advisable to combine such platform-dependent datasets to take advantage of training data developed for other platforms? Should such data just be added to the training set or they should be selected in some way? And what happens if training data are available only for one platform and not for the other?

In this paper we address all the above questions focusing on hate speech detection for Italian. After identifying a modular neural architecture that is rather stable and well-performing across different languages and platforms (Corazza et al., to appear), we perform our comparative evaluation on freely available datasets for hate speech detection in Italian, extracted from four different social media platform, i.e. Facebook, Twitter, Instagram and Whatsapp. In particular, we test the same model while altering only some features and pre-processing aspects. Besides, we use a multi-platform training set but test on data taken from the single platforms. We show that the proposed solution of combining platform-dependent datasets in the training phase is beneficial for all platforms but Twitter, for which results obtained by training on tweets only outperform those obtained with a training on the mixed dataset.

2 Related work

In 2018, the first *Hate Speech Detection* (HaSpeeDe) task for Italian (Bosco et al., 2018) has been organized at EVALITA-2018[2], the evaluation campaign for NLP and speech processing tools for Italian. The task consists in automatically annotating messages from Twitter and Facebook, with a boolean value indicating the presence (or not) of hate speech. Two cross-platform tasks (Cross-HaSpeeDe) were also proposed, where the training was done on platform-specific data (Facebook or Twitter) and the test on data from another platform (Twitter or Facebook). In general, as expected, results obtained for Cross-HaSpeeDe were lower compared to those obtained for the in-domain tasks, due to the heterogeneous nature of the datasets provided for the task, both in terms of class distribution and data composition. Indeed, not only are Facebook posts in the task dataset longer, but they are also on average more likely to contain hate speech (68% hate posts in the Facebook test set vs. 32% in the Twitter one). This led to a performance drop, with the best system scoring 0.8288 F1 on in-domain Facebook data, and 0.6068 when the same model is tested on Twitter data (Cimino et al., 2018).

The best performing systems on the cross-tasks were ItaNLP (Cimino et al., 2018) when training on Twitter data and testing on Facebook, and Inria-FBK (Corazza et al., 2018) in the other configuration. The former adopts a newly-introduced approach based on a 2-layer BiLSTM which exploits multi-task learning with additional data from the 2016 SENTIPOLC task[3]. The latter, instead, uses a simple recurrent model with one hidden layer of size 500, a GRU of size 200 and no dropout.

The Cross-HaSpeeDe tasks and the analysis of system performance in a cross-platform scenario are the starting point of this study. The task summary presented in (Bosco et al., 2018) listed some remarks on the elements affecting the system robustness that led us extend the cross-platform experiments to new platforms, including also WhatsApp and Instagram data. To our knowledge, there have not been attempts to develop Italian systems for hate speech detection on these two platforms, probably because of the lack of suitable datasets. We therefore annotate our own Instagram data for the task, while we take advantage of a recently developed dataset for cyberbullying detection to test our system on WhastApp.

3 Data and linguistic resources

In the following, we present the datasets used to train and test our system and their annotations (Section 3.1). Then, we describe the word embeddings (Section 3.2) we have used in our experiments.

3.1 Datasets

Twitter dataset released for the HaSpeeDe (Hate Speech Detection) shared task organized at EVALITA 2018. This dataset includes a total amount of 4,000 tweets (2,704 negative and 1,296 positive instances, i.e. containing hate speech), comprising for each tweet the respective annotation, as can be seen in Example 1. The two classes considered in the annotation are “hateful post” or “not”.

1. Annotation: hateful.
   altro che profughi? sono zavorre e tutti uomini (EN: other than refugees? they are ballast and all men).

Facebook dataset also released for the HaSpeeDe (Hate Speech Detection) shared task. It consists of 4,000 Facebook comments collected from 99 posts crawled from web pages (1,941 negative, and 2,059 positive instances), comprising for each comment the respective annotation, as can be seen in Example 2. The two classes considered in the annotation are “hateful post” or “not”.

2. Annotation: hateful.
   Matteo serve un colpo di stato. Qua tra poco dovremo andare in giro tutti armati come in America. (EN: Matteo, we need a coup. Soon we will have to go around armed as in the U.S.).

[2] http://www.evalita.it/2018

[3] http://www.di.unito.it/~tutreeb/sentipolc-evalita16/index.html

Whatsapp dataset collected to study pre-teen cyberbullying (Sprugnoli et al., 2018). Such dataset has been collected through a WhatsApp experimentation with Italian lower secondary school students and contains 10 chats, subsequently annotated according to different dimensions as the roles of the participants (e.g. bully, victim) and the presence of cyberbullying expressions in the message, distinguished between different classes of insults, discrimination, sexual talk and aggressive statements. The annotation is carried out at token level. To create additional training instances for our model, we join subsequent sentences of the same author (to avoid cases in which the user writes one word per message) resulting in 1,640 messages (595 positive instances). We consider as positive instances of hate speech the ones in which at least one token was annotated as a cyberbullying expression, as in Example 3).

3. Annotation: Cyberbulling expression.
   fai schifo, ciccione! (EN: you suck, fat guy).

Instagram dataset includes a total amount of 6,710 messages, which we randomly collected from Instagram focusing on students' profiles (6,510 negative and 200 positive instances) identified through the monitoring system described in (Menini et al., 2019). Since no Instagram datasets in Italian were available, and we wanted to include this platform to our study, we manually annotated them as "hateful post" (as in Example 4) or "not".

4. Annotation: hateful.
   Sei una troglodita (EN: you are a caveman).

3.2 Word Embeddings

In our experiments we test two types of embeddings, with the goal to compare generic with social media-specific ones. In both cases, we rely on Faxttext embeddings (Bojanowski et al., 2017), since they include both word and subword information, tackling the issue of out-of-vocabulary words, which are very common in social media data:

- **Generic embeddings**: we use embedding spaces obtained directly from the Fasttext website[4] for Italian. In particular, we use the Italian embeddings trained on Common Crawl and Wikipedia (Grave et al., 2018) with size 300. A binary Fasttext model is also available and was therefore used;
- **Domain-specific embeddings**: we trained Fasttext embeddings from a sample of Italian tweets (Basile and Nissim, 2013), with embedding size of 300. We used the binary version of the model.

4 System Description

Since our goal is to compare the effect of various features, word embeddings, pre-processing techniques on hate speech detection applied to different platforms, we use a modular neural architecture for binary classification that is able to support both word-level and message-level features. The components are chosen to support the processing of social-media specific language.

4.1 Modular neural architecture

We use a modular neural architecture (see Figure 1) in Keras (Chollet and others, 2015). The architecture that constitutes the base for all the different models uses a single feed forward hidden layer of 500 neurons, with a ReLu activation and a single output with a sigmoid activation. The loss used to train the model is binary cross-entropy. We choose this particular architecture because it showed good performance in the EVALITA shared task for cross-platform hate speech detection, as well as in other hate speech detection tasks for German and English (Corazza et al., to appear). The architecture is built to support both word-level (i.e. embeddings) and message-level features. In particular, we use a recurrent layer to learn an encoding (x_n in the Figure) derived from word embeddings, obtained as the output of the recurrent layer at the last timestep. This encoding gets then concatenated with the other selected features, obtaining a vector of message-level features.

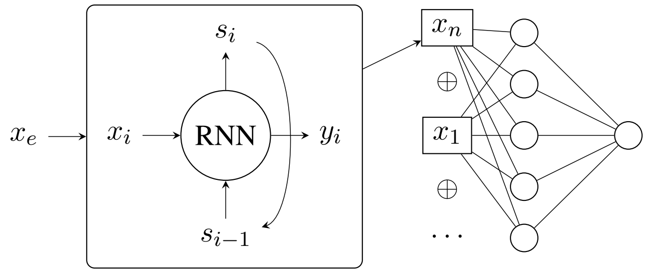

Figure 1: Modular neural architecture for Italian hate speech detection

[4] urlhttps://fasttext.cc/docs/en/crawl-vectors.html

4.2 Preprocessing

The language used in social media platforms has some peculiarities with respect to standard language, as for example the presence of URLs, ”@” user mentions, emojis and hashtags. We therefore run the following pre-processing steps:

- URL and mention replacement: both urls and mentions are replaced by the strings ”URL” and ”username” respectively;

- Hashtag splitting: Since hashtags often provide important semantic content, we wanted to test how splitting them into single words would impact on the performance of the classifier. To this end, we use the Ekphrasis tool (Baziotis et al., 2017) to do hashtag splitting and evaluate the classifier performance with and without splitting. Since the aforementioned tool only supports English, it has been adapted to Italian by using language-specific Google ngrams.[5]

4.3 Features

- **Word Embeddings**: We evaluate the contribution of word embeddings extracted from social media data, compared with the performance obtained using generic embedding spaces, as described in Section 3.2.

- **Emoji transcription**: We evaluate the impact of keeping emojis or transcribing them in plain text. To this purpose, we use the official plaintext descriptions of the emojis (from the unicode consortium website), translated to Italian with Google translate and then manually corrected, as a substitute for emojis

- **Hurtlex**: We assess the impact of using a lexicon of hurtful words (Bassignana et al., 2018), created starting from the Italian hate lexicon developed by the linguist Tullio De Mauro, organized in 17 categories. This is used to associate to the messages a score for ‘hurtfulness’

- **Social media specific features**: We consider a number of metrics related to the language used in social media platforms. In particular, we measure the number of hashtags and mentions, the number of exclamation and question marks, the number of emojis, the number of words written in uppercase

5 Experimental Setup

In order to be able to compare the results obtained while experimenting with different training datasets and features, we used fixed hyperparameters, derived from our best submission at EVALITA 2018 for the cross-platform task that involved training on Facebook data and testing on Twitter. In particular, we used a GRU (Cho et al., 2014) of size 200 as the recurrent layer and we applied no dropout to the feed-forward layer. Additionally, we used the provided test set for the two Evalita tasks, using 20% of the development set for validation. For Instagram and WhatsApp, since no standard test set is available, we split the whole dataset using 60% of it for training, while the remaining 40% is split in half and used for validation and testing. For this purpose, we use the *train_test_split* function provided by sklearn (Pedregosa et al., 2011), using 42 as seed for the random number generator.

One of our goals was to establish whether merging data from multiple social media platforms can be used to improve performance on single platform test sets. In particular, we used the following datasets for training:

- **Multi-platform**: we merge all the datasets mentioned in Section 3 for training.

- **Multi-platform filtered by length**: we use the same datasets mentioned before, but only considered instances with a length lower or equal to 280 characters, ignoring URLs and user mentions. This was done to match Twitter length restrictions.

- **Same Platform**: for each of the datasets, we trained and tested the model on data from the same platform.

In addition to the experiments performed on different datasets, we also compare the system performance obtained by using different embeddings. In particular, we train the system by using Italian Fasttext word embeddings trained on Common-Crawl and Wikipedia, and Fasttext word embeddings trained by us on a sample of Italian tweets

[5] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

Platform	Training set	Embeddings	Features	Emoji Transcription	F1 no hate	F1 hate	Macro AVG
Instagram	Multi Platform	Twitter	Social	Yes	0.984	0.432	0.708
	Single Platform	Twitter	Social	Yes	0.981	0.424	0.702
Facebook	Multi Platform	Twitter	Social	Yes	0.773	0.871	0.822
	Single Platform	Twitter	Social	Yes	0.733	0.892	0.812
WhatsApp	Multi Platform	Twitter	Social	Yes	0.852	0.739	0.796
	Single Platform	Twitter	Social	Yes	0.814	0.694	0.754
Twitter	Single Platform	Twitter	Hurtlex	No	0.879	0.717	0.798
	Filtered Multi Platform	Twitter	Hurtlex	No	0.858	0.720	0.789
	Multi Platform	Twitter	Hurtlex	No	0.851	0.712	0.782

Table 1: Classification results

(Basile and Nissim, 2013), with an embedding size of 300. As described in Section 4.3, we also train our models including either social-media or Hurtlex features. Finally, we compare classification performance with and without emoji transcription.

6 Results

For each platform, we report in Table 1 the best performing configuration considering embedding type, features and emoji transcription. We also report the performance obtained by merging all training data (*Multi-platform*), using only platform-specific training data (*Single platform*) and filtering training instances $>$ 280 characters (*Filtered Multi platform*) when testing on Twitter.

For Instagram, Facebook and Whatsapp, the best performing configuration is identical. They all use emoji transcription, Twitter embeddings and social-specific features. Using multi-platform training data is also helpful, and all the best performing models on the aforementioned datasets use data obtained from multiple sources. However, the only substantial improvement can be observed in the WhatsApp dataset, probably because it is the smallest one, and the classifier benefits from more training data.

The results obtained on the Twitter test set differ from the aforementioned ones in several ways. First of all, the in-domain training set is the best performing one, while the restricted length dataset is slightly better than the non restricted one. These results suggest that learning to detect hate speech on the short length interactions that happen on Twitter does not benefit from using data from other platforms. This effect can be at least partially mitigated by restricting the length of the social interactions considered and retaining only the training instances that are more similar to Twitter ones.

Another remark concerning only Twitter is that Hurtlex is in this case more useful than social network specific features. While the precise cause for this would require more investigation, one possible explanation is the fact that Twitter is known for having a relatively lenient approach to content moderation. This would let more hurtful words slip in, increasing the effectiveness of Hurtlex as a feature, in addition to word embeddings. Additionally, emoji transcription seems to be less useful for Twitter than for other platforms. This might be explained with the fact that the Twitter dataset has relatively less emojis when compared to the others.

One final outtake confirmed by the results is the fact that embeddings trained on social media platforms (in this case Twitter) always outperform general-purpose embeddings. This shows that the language used on social platforms has peculiarities that might not be present in generic corpora, and that it is therefore advisable to use domain-specific resources.

7 Conclusions

In this paper, we examined the impact of using datasets from multiple platforms in order to classify hate speech on social media. While the results of our experiments successfully demonstrated that using data from multiple sources helps the performance of our model in most cases, the resulting improvement is not always sizeable enough to be useful. Additionally, when dealing with tweets, using data from other social platforms slightly decreases performance, even when we filter the data to contain only short sequences of text. As for future work, further experiments could be performed, by testing all possible combinations of training sources and test sets. This way, we could establish what social platforms share more traits when it comes to hate speech, allowing for better detection systems. At the moment, however, the size of the datasets varies too broadly to allow for a fair comparison, and we would need to extend some of the datasets. Finally, another approach could be tested, where a model trained on Facebook is used for longer sequences of text, while the Twitter model is applied to the shorter ones.

Acknowledgments

Part of this work was funded by the CREEP project (`http://creep-project.eu/`), a Digital Wellbeing Activity supported by EIT Digital in 2018 and 2019. This research was also supported by the HATEMETER project (`http://hatemeter.eu/`) within the EU Rights, Equality and Citizenship Programme 2014-2020.

References

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. 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)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 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. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy.*

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

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. to appear. Robust Hate Speech Detection: A Cross-Language Evaluation. *Transactions on Internet Technology*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont. 2018. Proceedings of the 2nd workshop on abusive language online (alw2). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110, Florence, Italy, August. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59. Association for Computational Linguistics.

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

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.