# Source-driven Representations for Hate Speech Detection

**Flavio Merenda**[*∓], **Claudia Zaghi**[*], **Tommaso Caselli**[*], **Malvina Nissim**[*]
[*] Rikjuniversiteit Groningen, Groningen, The Netherlands
[∓] Università degli Studi di Salerno, Salerno, Italy
f.merenda|t.caselli|m.nissim@rug.nl c.zaghi@student.rug.nl

## Abstract

**English.** Sources, in the form of selected Facebook pages, can be used as indicators of hate-rich content. Polarized distributed representations created over such content prove superior to generic embeddings in the task of hate speech detection. The same content seems to carry a too weak signal to proxy silver labels in a distant supervised setting. However, this signal is stronger than gold labels which come from a different distribution, leading to re-think the process of annotation in the context of highly subjective judgments.

**Italiano.** *La provenienza di ciò che viene condiviso su Facebook costituisce un primo elemento indentificativo di contentuti carichi di odio. La rappresentazione distribuita polarizzata che costruiamo su tali contenuti si dimostra migliore nell'individuazione di argomenti di odio rispetto ad alternative più generiche. Il potere predittivo di tali embedding polarizzati risulta anche più incisivo rispetto a quello di dati gold standard che sono caratterizzati da una distribuzione ed una annotatione diverse.*

## 1 Introduction

Hate speech is "the use of aggressive, hatred or offensive language, targeting a specific group of people sharing a common trait: their gender, ethnic group, race, religion, sexual orientation, or disability" (Merriam-Webster's collegiate dictionary, 1999). The phenomenon is widely spread on-line, and Italian Social Media is definitely not an exception (Gagliardone et al., 2015). To monitor the problem, social networks and websites have introduced a stricter code of conduct and regularly remove hateful content flagged by users (Bleich, 2014). However, the volume of data requires that ways are found to classify on-line content automatically (Nobata et al., 2016; Kennedy et al., 2017).

The Italian NLP community is active on this front (Poletto et al., 2017; Del Vigna et al., 2017), with the development of labeled data, including the organization of a dedicated shared task at the EVALITA 2018 campaign[1]. Relying on manually labeled data has limitations, though: i.) annotation is time and resource consuming; ii.) portability to new domains is scarce[2]; iii.) biases are unavoidable in annotated data, especially in the form of annotation decisions. This is both due to the intrinsic subjectivity of the task itself, and to the fact that there is not, as yet, a shared set of definitions and guidelines across the different projects that yield annotated datasets.

Introduced as a new take on data annotation (Mintz et al., 2009; Go et al., 2009), *distant supervision* is used to automatically assign (silver) labels based on the presence or absence of specific hints, such as happy/sad emoticons (Go et al., 2009) to proxy positive/negative labels for sentiment analysis, Facebook reactions (Pool and Nissim, 2016; Basile et al., 2017) for emotion detection, or specific strings to assign gender (Emmery et al., 2017). Such an approach has the advantage of being more scalable (portability to different languages or domains) and versatile (time and resources needed to train), than pure supervised learning algorithms, while preserving competitive performance. Apart from the ease of generating labeled data, distant supervision has a valuable *ecological* aspect in not relying on third-party annotators to interpret the data (Purver and Battersby,

[1] http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html

[2] The EVALITA 2018 haspeede task addresses this issue by setting the task in a cross-genre fashion.

2012). This reduces the risk of adding extra bias (see also point (iii) about limitation in the previous paragraph), modulo the choices related to which proxies should be considered.

**Novelty and Contribution** We promote a special take on distant supervision where we use as proxies the *sources where the content is published on-line rather than any hint in the content itself.* Through a battery of experiments on hate speech detection in Italian we show that this approach yields meaningful representations and an increase in performance over the use of generic representations. Contextually, we show the limitations of silver labels, but also of gold labels that come from a different dataset with respect to the evaluation set.

## 2 Source-driven Representations

Our approach is based on previous studies on on-line communities showing that communities tend to reinforce themselves, enhancing "filter bubbles" effects, decreasing diversity, distorting information, and polarizing socio-political opinions (Pariser, 2011; Bozdag and van den Hoven, 2015; Seargeant and Tagg, 2018). Each community in the social media sphere thus represents a somewhat different source of data. Our hypothesis is that the contents generated by each community (source) can thus be used as proxies for specialized information or even labeled data.

Building on this principle, we scraped data from social media communities on Facebook, acquiring what we call *source-driven representations*. The data is indeed used in two ways in the context of Hate Speech detection, namely: i.) to generate (potentially) *polarized word embeddings* to be used in a variety of models, comparing it to more standard generic embeddings (Section 3); and ii.) as *training data* for a supervised machine learning classifier, combining and comparing it with manually labeled data (Section 4).

## 3 Polarized Embeddings

Polarized embeddings are representations built on a corpus which is not randomly representative of the Italian language, rather collected with a specific bias. In this context, we use data scraped from Facebook pages (communities) in order to create hate-rich embeddings.

**Data acquisition** We selected a set of publicly available Facebook pages that may promote or be the target of hate speech, such as pages known for promoting nationalism (*Italia Patria Mia*), controversies (*Dagospia*, *La Zanzara - Radio 24*), hate against migrants and other minorities (*La Fabbrica Del Degrado*, *Il Redpillatore*, *Cloroformio*), support for women and LGBT rights (*NON UNA DI MENO*, *LGBT News Italia*). Using the Facebook API, we downloaded the comments to posts as they are the text portions most likely to express hate, collecting a total of over 1M comments for almost 13M tokens (Table 1).

| **Page Name** | **Comments** |
|---|---:|
| Matteo Salvini | 318,585 |
| NON UNA DI MENO | 5,081 |
| LGBT News Italia | 10,296 |
| Italia Patria Mia | 4,495 |
| Dagospia | 41,382 |
| La Fabbrica Del Degrado | 6,437 |
| Boom. Friendzoned. | 85,132 |
| Cloroformio | 392,828 |
| Il Redpillatore | 6,291 |
| Sesso Droga e Pastorizia | 8,576 |
| PSDM | 44,242 |
| Cara, sei femminista - Returned | 830 |
| Se solo avrei studiato | 38,001 |
| La Zanzara - Radio 24 | 215,402 |
| **Total** | **1,177,578** |

Table 1: List of public pages from Facebook and number of extracted comments per page.

**Making Embeddings** We built distributed representations over the acquired data. The embeddings have been generated with the `word2vec`[3] skip-gram model (Mikolov et al., 2013) using 300 dimensions, a context window of 5, and minimum frequency 1. The final vocabulary amounts to 381,697 words.

These hate-rich embeddings are used in models for hate speech detection. For comparison, we also use larger, generic embeddings that were trained on the Italian Wikipedia (more than 300M tokens)[4] using `GloVe` (Berardi et al., 2015)[5]; the vocabulary amounts to 730,613 words. As a sanity check, and a sort of qualitative intrinsic evaluation, we probed our embeddings with a few keywords, reporting in Table 2 the top three nearest neighbors for the words "immigrati" [migrants]

[3] `https://radimrehurek.com/gensim/;https://github.com/RaRe-Technologies/gensim`

[4] `http://hlt.isti.cnr.it/wordembeddings/`

[5] `https://nlp.stanford.edu/projects/glove/`

and “trans”. For the former, it is interesting to see how the polarized embeddings return more hate-leaning words compared to the generic embeddings. For the latter, in addition to hateful epithets, we also see how these embeddings capture the correct semantic field, while the generic ones do not.

Table 2: Intrinsic embedding comparison: words most similar to potential hate targets.

| Generic Embeddings | Polarized Embeddings |
|---|---|
| “immigrati” [migrants] | |
| immigranti (0.737) | extracomunitari (0.841) |
| emigranti (0.731) | immigranti(0.828) |
| emigrati (0.725) | clandestini (0.823) |
| “trans” [trans] | |
| europ (0.399) | lesbo (0.720) |
| express (0.352) | puttane (0.709) |
| airlines (0.327) | gay (0.703) |

**Classification** To test the contribution of our embeddings, we used them in two different classifiers, comparing them to alternative distributed representations.

First, we built a Convolutional Neural Network (CNN), using the implementation of (Kim, 2014). This is a simple architecture with one convolutional layer built on top of a word embeddings layer (hyperparameters: `Number of filters`: 6; `Filter sizes`: 3, 5, 8; `Strides`: 1; `Activation function`: Rectifier). We experimented with three different activation strategies for the CNN model: i.) random initialization, by generating word embeddings from the training data itself, i.e. “on-the-fly”; ii.) pre-trained 300 dimension general word embeddings; iii.) our own polarised embeddings.

Second, and for further comparison, we also built a simple Linear Support Vector Machine (SVM), using the LinearSVC scikit learn implementation (Pedregosa et al., 2011). In one setting, we used only information coming from the two different sets of pre-trained embeddings (GloVe generic vs our polarized ones) to observe their contribution alone, in the same fashion as the CNN. To use these word vectors in the SVM model, we mapped the content words in each sentence and we replaced them with the corresponding word embeddings values; afterwards, we computed the average value for each word embedding, in order to achieve a unique one-dimensional sentence vector with each word replaced with the corresponding embedding average. In further settings, we combined this information with a more standard n-gram-based tf-idf model. Specifically, we use 1-3 word and 2-4 character n-grams, with default parameter values for the SVM.

We train and test our models using the manually labelled data provided in the context of the EVALITA 2018 task on Hate Speech Detection (`haspeede`) [6]. The released training/development set comprises 3000 Facebook comments and 3000 tweets. The proportion of hateful content in this dataset is 39%, with 46% in the Facebook portion, and 32% in Twitter. We train on 80% of `haspeede` (4800 instances), and test on the remaining 20%. We report precision, recall, and F-score per class, averaged over ten random train/test splits. To assess general performance, we use macro F-score rather than micro F-score as the classifier's accuracy on the minority class is particularly important. This is also reported as the average of the ten different runs.

**Results** The results in Table 3 show that despite our embeddings being almost 25 times smaller than the generic ones, they yield a substantially better performance both in the CNN model and in the SVM classifier. In the former, they are also more informative than the representations obtained on-the-fly from the training data. In the latter, the contribution of embeddings in general appears though rather marginal on top of a more standard SVM model based on n-gram tf-idf information, and the difference according to which representation is used is not significant. Finally, it is interesting to note that the polarized embeddings cover 55% of the tokens in the training data (*vs.* only 45% of the generic ones, in spite of the substantial size difference between the two.

## 4 Silver labels

In a more standard distantly supervised setting, modulo proxing labels via sources rather than specific keywords/emojis, we also used the scraped text as training data directly. Because we approximate labels with sources, and we had collected data from supposedly hate-rich pages, for the current experimental settings we balanced the data by

[6] `http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html`

Table 3: Results for the contribution of different embeddings in CNN and SVM models. The models are trained and tested on 80/20 splits randomised ten times on manually labelled data. Results are reported as averages. We underline the best score for each set of experiments, and boldface the best score overall.

| MODEL | CLASS | P | R | F | MACRO F |
|---|---|---|---|---|---|
| EMBEDDINGS ALONE | | | | | |
| CNN on-the-fly embeds | non-H | .84 | .75 | .79 | .749 |
| | H | .77 | .65 | .70 | |
| CNN generic embeds | non-H | .80 | .86 | .83 | .760 |
| | H | .74 | .65 | .69 | |
| CNN polarised embeds | non-H | .82 | .88 | .85 | .786 |
| | H | .78 | .68 | .73 | |
| SVM generic embeds | non-H | .77 | .85 | .81 | .728 |
| | H | .71 | .60 | .65 | |
| SVM polarised embeds | non-H | .79 | .84 | .81 | .750 |
| | H | .72 | .66 | .69 | |
| N-GRAMS + EMBEDDINGS | | | | | |
| SVM tf-idf + generic embeds | non-H | .84 | .87 | .85 | .806 |
| | H | .78 | .74 | .76 | |
| SVM tf-idf + polarised embeds | non-H | .84 | .86 | .85 | **.807** |
| | H | .78 | .75 | .76 | |
| N-GRAMS ALONE | | | | | |
| SVM tf-idf | non-H | .83 | .87 | .85 | .802 |
| | H | .78 | .72 | .75 | |

scraping Facebook comments from an Italian news agency (i.e. ANSA), assuming it conveys neutral content rather than polarized.

As for the distribution of labels, we followed the proportion of the Facebook portion of the haspeede dataset (46% of hateful content, and the rest non-polarized). We proxy labels according to sources, and under the above presumed proportions, we selected a total of 100,000 comments.

For comparison, and in combination, we also used gold data. In addition to the previously mentioned 6000 instances from the haspeede task, we used the Turin dataset, a collection of 990 manually labelled tweets concerning the topic of immigration, religion and Roma[7] (Poletto et al., 2017; Poletto et al., 2018). The distribution of labels in this dataset differs from the EVALITA dataset, with only 160 (16%) hateful instances.

We trained an SVM classifier with the best settings as observed in Section 3 (tf-idf and and polarised embeddings) using different training sets, combining gold and silver data (see Table 4). For evaluation, we use the same settings as the experiments in Section 3, by picking a random test set out of the haspeede dataset ten times, and reporting averaged results.

Table 4: Evaluation on 1200 instances from haspeede (averaged over 10 randomly picked test sets), using train sets from different sources and combinations thereof. The haspeede and Turin sets have gold labels.

| TRAINSET | CLASS | P | R | F | MACRO F |
|---|---|---|---|---|---|
| 100K silver | non-H | .60 | .39 | .47 | .464 |
| | H | .38 | .59 | .46 | |
| 3600 haspeede | non-H | .85 | .86 | .85 | .807 |
| | H | .77 | .76 | .76 | |
| 3600 haspeede + 1000 silver | non-H | .83 | .85 | .84 | .792 |
| | H | .76 | .73 | .74 | |
| 3600 haspeede + 990 Turin | non-H | .81 | .86 | .83 | .777 |
| | H | .76 | .68 | .72 | |
| 3600 haspeede + 1200 haspeede | non-H | .85 | .86 | .85 | .814 |
| | H | .78 | .77 | .77 | |

**Results** From Table 4 we can make the following observations: (i) training on silver labels lets us detect hate speech better than a most-frequent-label baseline (macro F=.383); (ii) however, in this context, training on small amounts of gold data is substantially more accurate than training on large amounts of distantly supervised data (.807 vs .464); (iii) adding even small amounts of silver data to gold decreases performance (.792 vs .807)[8]; (iv) also adding more gold data decreases performance, *even more so than adding an equal amount of silver data*, if the manually labeled data comes from a different dataset (thus created with different guidelines, and in this case with a different hate/non-hate distribution). Performance goes up as expected when adding more data from the same dataset (.814 vs .807).

## 5 Conclusions

We exploited distant supervision to automatically obtain representations from Facebook-scraped content in two forms. First, we generated polarized, hate-rich distributed representations which proved superior to larger, generic embeddings when used both in a CNN and an SVM model for hate speech detection. Second, we used the scraped data as training material directly, proxing

[7] The Romani, Romany, or Roma are an ethnic group of traditionally itinerant people who originated in northern India and are nowadays subject to ethnic discrimination.

[8] We also experimented with adding progressively larger batches of silver data to gold (2K, 3K, 5K, etc.), but this yielded a steady decrease in performance.

labels (hate vs non-hate) with the sources where the data was coming from (Facebook pages). This did not prove as a successful alternative nor complementary strategy to using gold data, though performance above baseline indicates some signal is present. Importantly, though, our experiments also suggest that gold data is not better than silver data if it comes from a different dataset. This highlights a crucial aspect related to the creation of manually labeled datasets, especially in the highly subjective area of hate speech and affective computing in general, where different guidelines and different annotators clearly introduce large biases and discrepancies across datasets.

All considered, we believe that obtaining data in a distant, more ecological way should be further pursued and refined. How to better exploit the information that comes from polarized embeddings in combination with other features is also left to future work.

## Acknowledgments

The authors want to thank the EVALITA 2018 Hate Speech Detection (HaSpeeDe) task organizers for allowing us to use their datasets.

## References

Angelo Basile, Tommaso Caselli, and Malvina Nissim. 2017. Predicting Controversial News Using Facebook Reactions. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy.

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Erik Bleich. 2014. Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the usa and europe. *Journal of Ethnic and Migration Studies*, 40(2):283–300.

Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, 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), Venice, Italy, January 17-20, 2017*, pages 86–95.

Chris Emmery, Grzegorz Chrupała, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 50–55.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12).

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the First Workshop on Abusive Language Online*, pages 73–77.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

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.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *CEUR WORKSHOP PROCEEDINGS*, volume 2006, pages 1–6. CEUR-WS.

Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. *In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan, December. COLING 2016.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics.

Philip Seargeant and Caroline Tagg. 2018. Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum. *Discourse, Context & Media*.