Hate Speech Detection with Machine-Translated Data: The Role of Annotation Scheme, Class Imbalance and Undersampling

Camilla Casula
Fondazione Bruno Kessler
Trento, Italy
ccasula@fbk.eu

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

Abstract

While using machine-translated data for supervised training can alleviate data sparseness problems when dealing with less-resourced languages, it is important that the source data are not only correctly translated, but also follow the same annotation scheme and possibly class balance as the smaller dataset in the target language. We therefore present an evaluation of hate speech detection in Italian using machine-translated data from English and comparing three settings, in order to understand the impact of training size, class distribution and annotation scheme.[1]

1 Introduction

The task of detecting hate speech on social media has been attracting increasing attention due to the negative effects this phenomenon can have on online communities and society as a whole. The development of systems which can effectively detect hate speech has therefore become increasingly important for academics and tech companies alike.

One of the difficulties of producing accurate hate speech detection systems is the need for large, high-quality datasets, the creation of which is time and resource-consuming. English can count on the highest number of hate speech detection datasets, as well as the ones with the largest sizes, with up to 150k posts for a single dataset (Gomez et al., 2020). Other languages such as Italian, on the other hand, can count on fewer datasets which tend to be smaller (Vidgen and Derczynski, 2020). Given that machine learning methods are typically used for this task, the use of small datasets can lead to overfitting problems due to the lack of linguistic variation (Vidgen and Derczynski, 2020). One possible solution to alleviate data sparseness is the use of machine translated data from English to less resourced languages for training classifiers, exploiting the large amount of data available for English. This has already been used in the context of hate speech detection (Sohn and Lee, 2019; Casula et al., 2020) but results have not been consistent across languages.

An additional issue is the fact that there is no shared fixed definition within the NLP community of what type of language constitutes hate speech. Indeed, there are typically large differences among hate speech and abusive language datasets in terms of annotation frameworks and their applications in practice (Caselli et al., 2020). In addition to this, there can be large variations between datasets in terms of size and class balance. Possible issues affecting the behaviour of classifiers trained on machine-translated data, such as different class distribution in source and target language, or different annotation scheme, have not been analysed.

In order to fill this gap, we explore the impact of these differences between datasets when performing hate speech detection in Italian using machine-translated data from English. Our goal is to address the three following questions:

- What performance can we expect by using only machine translated data, given that translation quality for social media language may be problematic?
- Is it better to use a larger translated set for training, even by merging slightly different classes, or a smaller, more precise one?
- What is the impact of class imbalance, and to what extent can undersampling be effective?

The above questions are addressed by comparing three experimental settings that are described in Section 4 and evaluated in Section 5.

2 Related Work

In recent years, the number of research works focused on the detection of hate speech on social media has remarkably increased, mostly due to the growing awareness regarding the societal impact these platforms can have.

Computational methods for detecting the presence of hate speech on the web have become necessary due to the extremely large amounts of user-generated content being posted each day. These methods typically rely on supervised learning, in the form of both traditional machine learning (e.g. support vector classifiers) and deep learning approaches (Schmidt and Wiegand, 2017). Given the increased attention towards this topic, more and more shared tasks regarding hate speech and abusive language detection have emerged, such as the HaSpeeDe task at Evalita 2018 (Bosco et al., 2018), OffensEval (Zampieri et al., 2019) and HatEval (Basile et al., 2019) at SemEval 2019, and the multilingual OffensEval at SemEval 2020 (Zampieri et al., 2020).

Systems based on Transformers architectures such as BERT (Devlin et al., 2019) have proven effective for hate speech detection and classification in both English (Zampieri et al., 2019) and Italian (Polignano et al., 2019a). These systems are generally pre-trained on large unlabeled corpora through two self-supervised tasks (next sentence prediction and masked language modeling) to create language models which can then be fine-tuned to a variety of downstream tasks using labeled data.

AlBERTo (Polignano et al., 2019b) is a BERT-based system which was pre-trained on Italian Twitter data, and it currently defines the state of the art for hate speech detection in Italian (Polignano et al., 2019a).

Recently, more attention has been directed towards the quality of hate and abuse detection systems. Vidgen et al. (2019) investigate the flaws presented by most abusive language detection datasets in circulation: they can contain systematic biases towards certain types and targets of abuse, they are subject to degradation over time, they typically present very low inter-annotator agreement, and they can vary greatly with respect to quality, size, and class balance. Vidgen and Derczynski (2020) further analyse the role of datasets in the detection of abuse, addressing issues such as the use of different task descriptions and annotation schemes across corpora, as well as similar annotation schemes being applied in different ways.

3 Data

Since tweets containing hate speech or abusive language constitute a very small subset (between 0.1% and 3% depending on the label used) of all tweets being posted (Founta et al., 2018), random samples are generally not used for annotation, because the final datasets would contain an extremely low number of positive class examples, which would make classification difficult. The typical solution to this is to preselect posts that are likely to contain hateful language by searching for specific hate-related keywords. While this method is effective for gathering more instances of hate speech, it can make datasets biased, which is a main issue in hate speech datasets (Wiegand et al., 2019).

The dataset we chose for training our system is described in Founta et al. (2018). This dataset was not created starting from a set of predefined offensive terms or hashtags in order to reduce bias, which was an important factor in our choice. The method used by Founta et al. (2018) to increase the percentage of hateful/abusive tweets is boosted random sampling, in which a portion of the dataset is "boosted" with tweets that are more likely to belong in the minority classes. The boosted set of tweets is created using text analysis and machine learning (Founta et al., 2018).

The dataset was annotated through crowdsourcing using the labels *hateful*, *abusive*, *spam*, and *normal*. The definition of *hate speech* given by Founta et al. (2018) to the annotators, based on existing literature on the topic, is:

> **Hate Speech**: Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.

The *abusive* label, on the other hand, is the result of three separate labels (*abusive*, *offensive*, and *aggressive*) being combined. In preliminary annotation rounds, Founta et al. (2018) found that these three labels were significantly correlated, so they grouped them together. The definition of *abusive language* given to the annotators is:

> **Abusive Language:** Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion.

While the Founta et al. (2018) dataset was originally comprised of 80k tweets, Twitter datasets can often be subject to degradation due to tweets being removed over time and not accessible anymore through tweet IDs (Vidgen et al., 2019). After retrieving all available tweets and after removing tweets annotated as spam, the total number of tweets we use for training is 12,379, of which 727 are annotated as hateful and 1,792 as abusive. Before translating the data into Italian, we preprocess it using the Ekphrasis tool [2] to tokenise the text and normalise user mentions, URLs (replaced by `<user>` and `<url>` respectively), as well as numbers, which are substituted with a `number` tag. We then use the Google Translate API to translate the data into Italian, in order to use it as training data for our classifier.

For testing, we use the test portion of the Twitter dataset used in the Hate Speech Detection (HaSpeeDe) task at Evalita 2018 (Bosco et al., 2018), consisting of 1,000 Italian tweets manually annotated for hate speech against immigrants. This dataset is a simplified version of the dataset described in (Sanguinetti et al., 2018), in which more fine-grained labels are used.

4 Experimental Setup

We experiment with the fine-tuning of AlBERTo (Polignano et al., 2019b), a BERT-based language model pre-trained on Italian Twitter data, using data that was automatically translated from English. This model has achieved state-of-the-art results when fine-tuned on the training data from the HaSpeeDe task at Evalita 2018 (Polignano et al., 2019a).

Our goal is that of exploring the impact of different annotation schemes and class balance when using machine-translated data for hate speech detection. Indeed, merging fine-grained classes into coarser ones has been a common and accepted practice when creating larger training sets from a smaller one (e.g. Founta et al. (2019)). This step has been performed also to compare classification in different languages (Corazza et al., 2020).

[2] https://github.com/cbaziotis/ekphrasis

In order to investigate this, we compare three different experimental settings. In the first one, we fine-tune AlBERTo on the translated tweets in Founta et al. (2018) after merging the *hateful* and *abusive* classes together, mapping them to a single hateful class as required by the binary classification task at Evalita 2018. In a second setting, AlBERTo is fine-tuned on the *hateful* class alone, discarding all tweets annotated as *abusive* in Founta et al. (2018). We hypothesize this setting may perform better when tested on the HaSpeeDe data, given the higher similarity in annotation framework.

Simply removing tweets annotated as abusive, however, can throw off the balance between classes. More specifically, when training the system on both abusive and hateful tweets the hateful+abusive class constitutes about 20% of our data, while when we only use tweets annotated as hateful this percentage drops to 7%, potentially affecting classification results. In particular, the data we use for testing has a different class balance, with 30% of tweets marked as hateful. In order to assess the impact of class imbalance on our results, we further evaluate each setting using undersampling (Kubat, 2000; Sun et al., 2009), a technique typically used for imbalanced classification, in which we reduce the number of tweets belonging to the majority class, so that the overall percentage of tweets containing hate increases.

Given that undersampling our data reduces the total size of tweets available for training, the resulting datasets for each annotation scheme considerably differ in size. We therefore consider a third setting, in which we use further random undersampling (Kubat, 2000; Sun et al., 2009) to match the larger dataset (hateful+abusive) with the smaller one (hateful only), so that the two annotations can be effectively compared in a setting with equal class balance and sample size.

In summary, the three data settings we train our system on are:

1. Hateful and abusive tweets, using undersampling to progressively lower class imbalance;

2. Hateful only tweets, again using undersampling to progressively lower class imbalance;

3. Hateful and abusive tweets, both using undersampling to progressively lower class imbalance as in the previous settings, and using

further random undersampling to match the low sample sizes of setting 2.

Our AlBERTo fine-tuning architecture consists of a pooling layer for extracting the AlBERTo hidden representation for each sequence, followed by a dropout layer (dropout rate 0.2), two dense layers of size 768 and 128 and, finally, a softmax layer. We use L2 regularization (λ=0.01), Adam optimizer (2e-5 learning rate), and categorical cross-entropy loss. We train the system for 5 epochs with batch size 32.

5 Results and Discussion

We measure the classification results using both macro-F1 score and minority class F1 score. We repeat each run five times in order to compensate for random initialization, and we report the average scores of these runs.

5.1 Setting 1: Hateful + Abusive Tweets

The classification results obtained when fine-tuning AlBERTo on both abusive and hateful tweets combined can be observed in Table 1.

Setting 1: Hateful + abusive

% hate	Size (tweets)	Macro-F1	Hate class F1
20%	12,379	0.40	0
30%	8,397	**0.64**	0.52
40%	6,298	0.63	**0.57**

Table 1: Scores obtained when fine-tuning AlBERTo on both hateful and abusive tweets.

The class balance of the dataset prior to undersampling is 20% hateful + abusive tweets and 80% non-hateful, which amounts to 12,379 tweets total. With this class balance, the system performs the worst, classifying every tweet as belonging to the majority non-hateful class. On the other hand, with a higher percentage of minority class instances, the classification results improve, in spite of the considerably smaller amount of training data available. These results suggest that consistency in class balance can play a bigger role than training data size in classification results in this context.

5.2 Setting 2: Hateful Only Tweets

The performance of the system when fine-tuned on tweets labeled as hateful only is reported in Table 2. As previously mentioned, only 7% of tweets in the dataset we use are labeled as hateful. The classes are therefore extremely imbalanced before undersampling. Predictably, with the classes being this imbalanced, the system identifies all test instances as belonging to the majority class. This again happens with the minority class comprising 20% of the training data.

Setting 2: Hateful only

% hate	Size (tweets)	Macro-F1	Hate class F1
7%	10,587	0.40	0
20%	3,635	0.40	0
30%	2,423	**0.65**	0.54
40%	1,818	0.52	**0.56**

Table 2: Scores obtained when fine-tuning AlBERTo on tweets labeled as hateful only.

Similarly to Setting 1, the best classification performance in this case is achieved with 30% of minority class tweets. Interestingly, the best performance is comparable to the one obtained in Setting 1, even though in this case the number of training samples available is much lower, suggesting that more task-specific training instances can impact performance. We can note a difference with the minority class at 40% of total data, in which the performance drops in terms of macro-F1 score, likely due to the very small number of samples available for training and the consequent lack of linguistic variation. The hate class F1 score, however, remains stable.

State-of-the-art results obtained by fine-tuning AlBERTo on the same Evalita dataset as reported in Polignano et al. (2019a) reach 0.80 macro-F1 and 0.73 F1 on the hate class, which we can consider an upper-bound for our task, obtained in a fully-supervised monolingual setting. On the other hand, the most frequent label baseline is 0.40 macro-F1, which is clearly outperformed using only machine-translated data.

5.3 Setting 3: Hateful + Abusive Tweets (Random Undersampling)

Since there are large differences in size between the hateful+abusive annotation and the hateful-only annotation, we randomly undersample the hateful+abusive training data so that it matches the size of the hateful-only training data, in order to allow us to effectively compare the impact of each annotation framework on our results. The classification performance is reported in Table 3.

If we compare the results of Setting 3 with those of Setting 2, it is clear that using more task-

Setting 3: Hateful + abusive (random undersampling)

% hate	Size (tweets)	Macro-F1	Hate class F1
30%	2,423	0.58	0.38
40%	1,818	**0.59**	**0.51**

Table 3: Scores obtained when fine-tuning AlBERTo on tweets labeled as hateful and abusive, after random undersampling.

specific data, in this case hateful-only tweets, can lead to a larger improvement in performance when the amount of training data is the same. This suggests that consistency in annotation between training and test data can have a positive impact on classification, although it is not fundamental to help classification of hate speech detection with machine translated data. In fact, other aspects such as class balance can also play an important role.

5.4 Qualitative Analysis

Another aspect affecting classification, which we have not considered so far, is the quality of machine translation, a particularly challenging task on social media data (Michel and Neubig, 2018). In order to assess the impact of translation quality on our results, two annotators with linguistic background manually analysed 500 samples from the training data, consisting of 300 tweets annotated as normal, 100 as hateful, and 100 as abusive. Each annotator checked manually 250 random tweets from this sample. Translation quality was evaluated using the semantic adequacy annotation scheme proposed in Dorr et al. (2011, p. 807). Annotations are judged on a scale between -3 and 3, with scores below 0 for inadequate translations and above 0 for adequate ones. The averaged annotations for each class are reported in Table 4.

	Normal	Hateful	Abusive	Overall
Average	0.438	0.527	-0.043	0.368

Table 4: Average translation quality scores.

Overall, translations tend towards adequacy, but the average scores are below 1 for all classes. Interestingly, tweets annotated as abusive show poorer translation quality than other classes. This could help explain the small differences in classification performance between our experiments.

A major role is played in this context by profanities, which are often used to offend a target but can also appear in non derogatory messages exchanged among members of the same community (Pamungkas et al., 2020). In the case of abusive tweets, we observe that the offenses are less direct and therefore slurs tend to be translated poorly. See for example the following sentence, which is labeled as abusive in the Founta et al. (2018) dataset:

(1) use that ugly ass design [...]
utilizzare quel disegno asino brutto [...]
use that design donkey ugly [...]

Here, "ass" is translated with "asino" ("donkey"), effectively removing the profanity in the translated tweet and changing completely the meaning of the message.

On the other hand, when profanities are used in a more direct way, or when they are expressed through unambiguous words such as "idiot" and "stupid", they tend to be translated correctly, contributing to a correct classification. Example 2 shows a hateful tweet which was translated almost correctly, retaining its offensiveness in the target language.

(2) what happens when you put idiots in charge
cosa succede quando si mette idioti in carica

6 Conclusions

In this paper we analysed the impact of machine-translated data on Italian hate speech detection in a zero-shot setting. Our experiments show that when using machine-translated data for training it is possible to learn a classification model that clearly outperforms the most-frequent baseline, even if translation quality is affected by the jargon used in social media data. We found that using more task-specific data can have a positive impact on classification performance even with lower sample sizes compared to larger, less targeted datasets.

Consistency in class distribution of training and test data can have a bigger impact than the size of the training set, or the annotation scheme. Indeed, using only the original training set translated into Italian, without undersampling, classification performance would be poor.

In the future, we plan to extend this kind of evaluation to new language pairs and new datasets, to check whether the findings obtained on the English – Italian pair are confirmed also with other languages.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9, Turin, Italy. CEUR.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6193–6202. European Language Resources Association.

Camilla Casula, Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Fbk-dh at semeval-2020 task 12: Using multi-channel bert for multilingual offensive language detection. In *Proceedings of Offenseval*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bonnie J Dorr, Joseph Olive, John McCary, and Caitlin Christianson, 2011. *Machine Translation Evaluation and Optimization*, pages 745 – 843. Springer New York.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *12th International AAAI Conference on Web and Social Media*.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 105–114, New York, NY, USA. Association for Computing Machinery.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, 03.

M. Kubat. 2000. Addressing the curse of imbalanced training sets: One-sided selection. *Fourteenth International Conference on Machine Learning*, 06.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium, October-November. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6237–6246. European Language Resources Association.

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

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019b. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International*

Workshop on Natural Language Processing for Social Media, pages 1–10, Valencia, Spain, April. Association for Computational Linguistics.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.

Y. Sun, A. Wong, and M. Kamel. 2009. Classification of imbalanced data: a review. *Int. J. Pattern Recognit. Artif. Intell.*, 23:687–719.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *ArXiv*, abs/2004.01670.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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