Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-20)

Modeling Annotator Perspective and Polarized Opinions to Improve Hate Speech Detection

Sohail Akhtar, Valerio Basile, Viviana Patti
Computer Science Department
University of Turin, Italy
sohail.akhtar@edu.unito.it, {basile, patti}@di.unito.it

Abstract

In this paper we propose an approach to exploit the fine-grained knowledge expressed by individual human annotators during a hate speech (HS) detection task, before the aggregation of single judgments in a gold standard dataset eliminates non-majority perspectives. We automatically divide the annotators into groups, aiming at grouping them by similar personal characteristics (ethnicity, social background, culture etc.). To serve a multi-lingual perspective, we performed classification experiments on three different Twitter datasets in English and Italian languages. We created different gold standards, one for each group, and trained a state-of-the-art deep learning model on them, showing that supervised models informed by different perspectives on the target phenomena outperform a baseline represented by models trained on fully aggregated data. Finally, we implemented an ensemble approach that combines the single perspective-aware classifiers into an inclusive model. The results show that this strategy further improves the classification performance, especially with a significant boost in the recall of HS prediction.

Introduction

Hate Speech is a special type of abusive language. It has specific targets which are victimized based on their personal characteristics or demographic background such as race, ethnicity, religion, color, sexual orientation or other similar factors (Nobata et al. 2016). Researchers who recently started tackling hate speech detection from a natural language processing perspective are designing operational frameworks for HS, annotating corpora with several semantic frameworks, and automatic classifiers based on supervised machine learning models (Fortuna and Nunes 2018; Schmidt and Wiegand 2017; Poletto et al. 2020).

Most datasets for HS detection are annotated by humans, often relying on crowd-sourcing, whereas typically no background information about the workers is provided. Given the highly subjective nature of HS, such datasets tend to exhibit low agreement by traditional measures. Moreover, aggregation by majority makes it difficult to model the different perspective of the annotators.

We propose a methodology to automatically model the different perspectives that annotators may adopt towards certain highly subjective phenomena, i.e., abusive language and hate speech. In our method, supervised machine learning models are trained to learn different points of view of the human annotators on the same data, in order to subsequently take them into account at prediction time. In this study, we will try to answer the following research questions: *(RQ1) Does an automatic partition of the annotators based on the polarization of their judgments reflect different perspectives on hate speech? (RQ2) Are models trained to represent such perspectives effective in HS detection tasks?*
In order to test these research questions, we experimented on three datasets in English and Italian.

Related Work

Hate Speech is a complex phenomenon which is dependent on the relationships between various communities and social groups. (Poletto et al. 2020) mention several definitions of hate speech, although there is no consensus on one formal definition (Ross et al. 2016). Therefore, it is difficult to develop automatic systems that determine whether a message contains any fragments of hate speech. A recent literature survey on hate speech detection (Fortuna and Nunes 2018) addresses many issues faced by researchers, including the scarcity of high quality datasets available as benchmarks for the hate speech detection tasks.

There are approaches that measure the level of controversy by analyzing user opinions on controversial topics. (Soberón et al. 2013) highlight the importance of disagreement in data annotations and treated it as a useful resource rather than noise in gold standard data. The majority of computational approaches to hate speech detection are based on supervised machine learning, including deep learning, but also Support Vector Machines, Random Forest, Logistic Regression and Decision Trees (Fortuna and Nunes 2018; Schmidt and Wiegand 2017). In particular, the state of the art is represented by deep learning models based on Transformer networks pre-trained on large amounts of unlabelled data and fine-tuned on task-specific annotated corpora.

Recently, neural language models have gained popularity. These models have been effectively applied to many NLP related tasks showing substantial improvements in the performances (Peters et al. 2018). Some of these pre-training based language models involve either feature-based approaches in which they only use pre-training as extra features, and depend on task-specific architectures such as ElMo (Peters et

al. 2018). (Howard and Ruder 2018) proposed the ULMFiT model for text classification tasks, achieving state-of-the-art performance on several benchmarks.

BERT (Devlin, Chang, and Toutanova 2019) is one of the best known Transformer-based models employing a bidirectional approach that achieved state of the art performance in many NLP tasks, in particular text classification (Yu, Jindian, and Luo 2019). BERT trains bidirectional language representations from unlabelled text and it considers both left and right contexts in a layered architecture (Munikar, Shakya, and Shrestha 2019).

Method

Our proposed method is based on the assumption that a group of annotators can be divided into groups based on some characteristics such as cultural background, common social behaviour and other similar factors. The idea is to investigate how these characteristics can influence the opinions of annotators expressed while annotating HS data. The method works in two steps, and it is applied to an annotated dataset for which the single, pre-aggregated annotations are known:

1. We divide the annotators into groups (two, in this iteration of the study) by using a numeric index measuring the polarization of the judgments.
2. Different gold standard datasets are compiled following the division of the annotators, and each used to train a different classifier.

The original and group-based models are tested against the same test set for comparison. The steps of the method are detailed in the rest of this section.

Division of the Annotators into Groups

The first step of our method consists in automatically dividing the set of annotators into groups. The group split is found by an exhaustive search of the possible annotator partitions and finding a partition which maximizes the average *Polarization index* (P-index). Such metric, introduced in (Akhtar, Basile, and Patti 2019), leverages the information at the single annotation level, measuring the level of polarization of all the annotations on each instance individually. The measurement of the P-index of a message is a three-step process. First, the annotators are divided into groups (in this study, we limit the possible partitions to two groups). Second, the agreement between the annotations of each group on each instance is measured by using the normalized χ^2 statistics, measuring how independent is their distribution compared to a uniform distribution. Finally, the P-index is computed as a function of the overall agreement and each of the per-group agreement values.

Once the annotator bi-partition is found that maximizes the average P-index, we create two new gold standard datasets, one for each individual group, by aggregating the annotations with a standard procedure of majority voting.

Supervised Classification

We employ the Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, and Toutanova 2019) as prediction framework for the binary classification task of hate speech detection.

In the pre-training step, the model is trained on different tasks on large unlabelled datasets. During the fine-tuning step, the pre-trained parameters are adjusted according to a specific task requirements and all the parameters are fine-tuned with labelled data from a downstream task.
In the second step of our method, *we fine-tune BERT models to the group-based gold standard datasets obtained in the previous steps, in order to learn different points of view on the perception of the same phenomenon (HS) on the same data.* By contrast, the model trained on the original dataset encodes all the possible points of view of the annotators.

Many BERT pre-trained models are available for multiple and individual languages, and trained on text from different genres and domains (Nozza, Bianchi, and Hovy 2020). In this work, we use the uncased base English model provided by Google for English (*uncased_L-12_H-768_A-12*). For Italian, we use AlBERTo (Polignano et al. 2019), a model for Italian, pre-trained on Twitter data. AlBERTo has similar specifications to the BERT English base model.

Data

We test our methodology on three data sets in English and Italian languages. The first and second datasets in English language are taken from previous work by (Waseem 2016). The original dataset contains 6,909 messages from Twitter annotated in a multi-label fashion with four labels: *sexism*, *racism*, *both*, and *neither*. We separated the corpus into two binary datasets, namely Sexism and Racism. We were able to retrieve a smaller dataset containing 6,361 tweets. There are 5,551 negative instances of HS and 810 positive in the Sexism dataset and 6,261 negative and 100 positive instances of HS in the Racism dataset. The third dataset is in Italian language, containing 1,859 tweets on topics related to LGBT community (1,635 negative and 224 positive instances).

We compiled the **Sexism** dataset as a binary classification dataset, mapping the labels *sexism* and *both* in the original dataset to the positive class, and the labels *racism* and *neither* to the negative class. The resulting Sexism dataset has 810 positive (sexist) tweets out of 6,361 (12.7%). The original dataset was annotated by experts (feminist and anti-racism activists) and workers on a crowd-sourcing platform[1]. The guidelines developed by (Waseem and Hovy 2016) were used to annotate the dataset. Majority voting was used to create a gold standard. After dividing the annotators in two by following the method introduced in the Method section, we report an overall agreement (Fleiss' Kappa among all annotators) of 0.58. The intra-group agreement (Cohen's Kappa only between the annotators of a group) for group one is 0.53 and 0.64 for group two.

We applied the same scheme to separate the **Racism** dataset from the original dataset that we applied for the Sexism dataset, except for the different labels. In particular, for the Racism dataset, *racism* and *both* were mapped to the positive class, whereas the labels *sexism* and *neither* were

[1] https://www.figure-eight.com/

mapped to the negative class. The final dataset comprises 100 positive (racist) tweets out of 6,361 (1.57%). We measured the overall agreement between all annotators and the value of Fleiss' Kappa is 0.23, which shows that there is a high disagreement between the annotators. We measured the intra-group agreement for the two groups as 0.22 and 0.25 respectively.

The **Homophobia** dataset in Italian language is an output of the ACCEPT European research project[2]. The dataset consists of tweets annotated with hate speech against the LGBT+ community. The original dataset was annotated in a multi-class fashion by five volunteers with four categories: *homophobic*, *not homophobic*, *doubtful* or *neutral*. We mapped *not-homophobic*, *doubtful* and *neutral* to the negative class (not homophobic) and the label *homophobic* to the positive class. These volunteers were hired by the main Italian non-for-profit organization for the LGBT+ rights Arcigay[3]. The annotators were selected to fill different demographic features such as age, education and personal view on LGBT+ stances to chose the volunteers for this important project. Some members of the LGBT+ community also volunteered to annotated the homophobia dataset. The overall agreement measured by using Fleiss Kappa is 0.35, rather low value according to common interpretation. The values of intra-group agreement for the two groups are 0.40 and 0.39 respectively.

Evaluation

The datasets presented in the Data section are employed to experiment with the method introduced in the Method section. For all datasets, the training set contains 80% of the dataset whereas, the remaining 20% constitutes the test set. We fine-tuned the BERT models on the training sets, keeping the test sets fixed for each dataset, for fair comparison. After a preliminary study, we fixed the sequence length at 128 words. The batch size was set to 12 for English and 8 for Italian, also due to memory limitations. The learning rate is $1e^{-5}$. We repeated each experiment five times, in order to average out the variance induced by the random initialization of the network.

The classification performance on the gold standard created by majority voting from the original datasets (before partition) are reported as baselines. We then test the performance of the two models trained on gold standard training sets created by only considering one group of annotators at a time (Group 1 and Group 2).

We also include the results obtained by a straightforward ensemble classifier which considers an instance positive if any of the Group 1 or Group 2 classifiers (or both) considers it positive. We call this ensemble "Inclusive". The rationale behind this ensemble is that hate speech is a sparse and subjective phenomenon, where each personal background induces a perspective that lead to different perceptions of what constitutes hate. This classifier includes all these perspectives in its decision process. The Inclusive classifier will naturally have a bias towards the positive class, by construction.

Tables 1, 2, and 3 show the results of the performed experiments. The results report the arithmetic mean of the evaluation metrics across five runs, along with their standard deviation, showing an improvement in our baseline on all datasets.

Table 1: Results of the prediction on the Sexism dataset. Averages of 5 runs with standard deviation in parenthesis.

Classifier	Prec. (1)	Rec (1)	F1 (1)
Baseline	**.812** (.034)	.711 (.044)	.756 (.015)
Group 1	.745 (.048)	.764 (.045)	.752 (.008)
Group 2	.720 (.019)	.907 (.018)	**.802** (.008)
Inclusive	.665 (.033)	**.939** (.009)	.778 (.020)

Table 2: Results of the prediction on the Racism dataset. Averages of 5 runs with standard deviation in parenthesis.

Classifier	Prec. (1)	Rec. (1)	F1 (1)
Baseline	**.852** (.159)	.194 (.059)	.312 (.085)
Group 1	.654 (.154)	.424 (.140)	.488 (.104)
Group 2	.571 (.175)	.412 (.198)	.419 (.076)
Inclusive	.532 (.141)	**.612** (.136)	**.542** (.091)

Table 3: Results of the prediction on the Homophobia dataset. Averages of 5 runs with standard deviation in parenthesis.

Classifier	Prec. (1)	Rec. (1)	F1 (1)
Baseline	.415 (.146)	.231 (.079)	.273 (.038)
Group 1	.302 (.038)	.471 (.154)	.355 (.040)
Group 2	**.531** (.112)	.178 (.031)	.262 (.033)
Inclusive	.302 (.039)	**.502** (.142)	**.367** (.035)

It is important to note that the improvement on the positive class is particularly important in this setting, since this binary classification task is actually a *detection* task. For the Sexism and Racism datasets, the overall improvement is mainly due to a better recall on the positive class. Precision drops less substantially, leading to better F1 scores. For the Homophobia dataset, group-based classifiers obtain an even greater improvement over the baseline, with higher precision, recall and F1 scores for the positive class.

The baseline results on the Racism and Homophobia datasets see substantially low recall values, which is expected given the highly skewed class distribution. Group-based classifiers largely correct this problem, although introducing some false positives (hence the lower precision on the positive class).

Finally, the results of the Inclusive ensemble classifier show that including multiple perspectives into the learning process is beneficial to the classification performance on all the datasets, however at the cost of lower precision.

[2]http://accept.arcigay.it/

[3]https://www.arcigay.it/en/

Conclusion and Future Work

In this paper, we presented a method to divide the annotators into groups based on their annotation behaviour, under the hypothesis that such partition reflects characteristics such as cultural background, common social behaviour and similar factors. We experimented with three social media datasets in English and Italian, reporting improvements over the baseline across all the datasets. The implementation of an "inclusive" classifier further boosts the classification performance by strongly increasing the recall on hateful messages.

Although the method boosts the hate speech classification performance, there are limitations which are important to consider. First, for the methodology to work, we need pre-aggregated data, which is often not available. Another issue is epistemological: our methodology and the subsequent empirical evaluation show that there is a great deal of information that is effectively wiped out by the aggregation step employed in the standard procedure to create benchmark datasets. This consideration motivates us to strongly promote the publication of datasets in pre-aggregated form, and to develop new paradigms of evaluation that take all the perspectives due to different backgrounds into account.

We plan to apply the methodology presented in this paper to other abusive language phenomena such as cyberbullying, radicalization, and extremism. We are also interested to test the method on sentiment analysis tasks applied to specific domains such as political debates.

We plan to investigate the effect of dividing the annotators into more than two groups, and how to find an optimal number of partitions. In this direction, unsupervised clustering of the annotators based on their annotations with standard methods (e.g. agglomerative) may be a solution both to the issue of the unavailability of background information on the annotators, and to the problem of computational complexity and scalability of the exhaustive search approach.

References

Akhtar, S.; Basile, V.; and Patti, V. 2019. A new measure of polarization in the annotation of hate speech. In Alviano, M.; Greco, G.; and Scarcello, F., eds., *AI*IA 2019 – Advances in Artificial Intelligence*, 588–603. Cham: Springer International Publishing.

Devlin, J.; Chang, M.-W.; and Toutanova, K. 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)*, 4171–4186. ACL.

Fortuna, P., and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys* 51:1–30.

Howard, J., and Ruder, S. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. ACL.

Munikar, M.; Shakya, S.; and Shrestha, A. 2019. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, 1–5. IEEE.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 145–153.

Nozza, D.; Bianchi, F.; and Hovy, D. 2020. What the [mask]? making sense of language-specific BERT models. *CoRR* abs/2003.02912.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 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)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics.

Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; and Patti, V. 2020. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*. To appear.

Polignano, M.; Basile, P.; de Gemmis, M.; Semeraro, G.; and Basile, V. 2019. 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 of *CEUR Workshop Proceedings*. Bari, Italy: CEUR-WS.org.

Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Beißwenger, M.; Wojatzki, M.; and Zesch, T., eds., *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, 6–9.

Schmidt, A., and Wiegand, M. 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*, 1–10. Valencia, Spain: Association for Computational Linguistics.

Soberón, G.; Aroyo, L.; Welty, C.; Inel, O.; Lin, H.; and Overmeen, M. 2013. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web – Volume 1030*, 45–58. CEUR-WS.org.

Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, 88–93. San Diego, California: ACL.

Waseem, Z. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. Austin, Texas: ACL.

Yu, S.; Jindian, S.; and Luo, D. 2019. Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access* PP:1–1.