Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter

Zeerak Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

Abstract

Hate speech in the form of racism and sexism is commonplace on the internet (Waseem and Hovy, 2016). For this reason, there has been both an academic and an industry interest in detection of hate speech. The volume of data to be reviewed for creating data sets encourages a use of crowd sourcing for the annotation efforts.

In this paper, we provide an examination of the influence of annotator knowledge of hate speech on classification models by comparing classification results obtained from training on expert and amateur annotations. We provide an evaluation on our own data set and run our models on the data set released by Waseem and Hovy (2016).

We find that amateur annotators are more likely than expert annotators to label items as hate speech, and that systems trained on expert annotations outperform systems trained on amateur annotations.

1 Introduction

Large amounts of hate speech on exists on platforms that allow for user generated documents, which creates a need to detect and filter it (Nobata et al., 2016), and to create data sets that contain hate speech and are annotated for the occurrence of hate speech. The need for corpus creation must be weighted against the psychological tax of being exposed to large amounts of abusive language (Chen, 2012).

A number of studies on profanity and hate speech detection, have crowdsourced their annotations due to the resources required to annotate large data sets and the possibility of distributing the load onto the crowd (Warner and Hirschberg, 2012; Nobata et al., 2016). Ross et al. (2016) investigate annotator reliability for hate speech annotation, concluding that *"hate speech is a fuzzy construct that requires significantly better definitions and guidelines in order to be annotated reliably"*.

Hate speech is hard to detect for humans (Sue et al., 2007), which warrants a thorough understanding of the benefits and pitfalls of crowdsourced annotation. This need is reinforced by previous studies, which utilize crowdsourcing of hate speech without knowledge on the quality of crowdsourced annotations for hate speech labeling.

In addition, it is important to understand how different manners of obtaining labeling can influence the classification models and how it is possible to obtain good annotations, while ensuring that annotators are not likely to experience adverse effects of annotating hate speech.

Our contribution We provide annotations of $6,909$ tweets for hate speech by annotators from CrowdFlower and annotators that have a theoretical and applied knowledge of hate speech, henceforth *amateur* and *expert* annotators[1]. Our data set extends the Waseem and Hovy (2016) data set by $4,033$ tweets. We also illustrate, how amateur and expert annotations influence classification efforts. Finally, we show the effects of allowing majority voting on classification and agreement between the amateur and expert annotators.

[1]Data set available at http://github.com/zeerakw/hatespeech

Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pages 138–142,
Austin, TX, November 5, 2016.

2 Data

Our data set is obtained by sampling tweets from the 130k tweets extracted by Waseem and Hovy (2016). The order of the tweets is selected by our database connection, thus allowing for an overlap with the data set released by Waseem and Hovy (2016). We find that there is an overlap of $2,876$ tweets (see Table 1) between the two data sets.

	Racism	Sexism	Neither
Count	1	95	2780

Table 1: Overlap between our data set and Waseem and Hovy (2016), denoted by their labels

Given the distribution of the labels in Waseem and Hovy (2016) and our annotated data set (see Table 2), it is to be expected the largest overlap occurs with tweets annotated as negative for hate speech. Observing Table 2, we see that the label distribution in our data set generally differs from the distribution in Waseem and Hovy (2016). In fact, we see that the amateur majority voted labels is the only distribution that tends towards a label distribution similar to Waseem and Hovy (2016), while the labels the amateurs fully agreed upon and the expert annotations have similar distributions.

	Racism	Sexism	Neither	Both
Expert	1.41%	13.08%	84.19%	0.70%
Amateur Majority	5.80%	19.00%	71.94%	1.50%
Amateur Full	0.69%	14.02%	85.15%	0.11%
Waseem and Hovy (2016)	11.6%	22.6%	68.3%	–

Table 2: Label distributions of the three annotation groups and Waseem and Hovy (2016).

Our annotation effort deviates from Waseem and Hovy (2016). In addition to "racism", "sexism", and "neither", we add the label "both" for tweets that contain both racism and sexism. We add this label, as the intersection of multiple oppressions can differ from the forms of oppression it consists of (Crenshaw, 1989), and as such becomes a unique form of oppression. Thus, we introduce a labeling scheme that follows an intersectional approach (Crenshaw, 1989). We do not require annotators to follow links. Instead, we ask them to annotate tweets only containing links as "Neither".

Expert Annotations We recruit feminist and anti-racism activists to annotate the data set. We present the annotators with the tests from Waseem and Hovy (2016). If a tweet fails any of the tests, the annotators are instructed to label it as the relevant form of hate speech. Expert annotators are given the choice of skipping tweets, if they are not confident in which label to assign, and a "Noise" label in case the annotators are presented with non-English tweets. Due to privacy concerns, all expert annotators are treated as a single entity.

Amateur Annotations Amateur annotators are recruited on CrowdFlower without any selection, to mitigate selection biases. They are presented with $6,909$ tweets that have been annotated by the expert annotators. The amateur annotators are not provided with the option to skip tweets, as they are not presented tweets the experts had skipped or labeled as "Noise".

Annotator agreement Considering annotator agreement, we find that the inter-annotator agreement among the amateur annotators is $\kappa = 0.57$ ($\sigma = 0.08$).

	Majority Vote	Full Agreement
Expert	0.34	0.70

Table 3: Kappa scores comparing majority voted label and full agreement with expert annotations.

The low agreement in Table 2 provides further evidence to the claim by Ross et al. (2016) that annotation of hate speech is a hard task. Table 2 suggests that if only cases of full agreement are considered, it is possible to obtain good annotations using crowdsourcing.

Overlap Considering the overlap with the Waseem and Hovy (2016), we see that the agreement is extremely low (mean pairwise $\kappa = 0.14$ between all annotator groups and Waseem and Hovy (2016)). Interestingly, we see that the vast majority of disagreements between our annotators and Waseem and Hovy (2016), are disagreements where our annotators do not find hate speech but Waseem and Hovy (2016) do.

3 Evaluation

We evaluate the influence of our features on the classification task using 5-fold cross validation to assess

		Amateur			Expert		
	Feature	F1	Recall	Precision	F1	Recall	Precision
Close	*Character n-gram*	86.41	88.53%	87.21%	91.24	92.49%	92.73%
	Token n-gram	86.37	88.60%	87.68%	91.55	92.92%	91.50%
	Token unigram	86.46	88.68%	87.74%	91.15	92.41%	92.37%
	Skip-grams	86.27	88.53%	87.62%	91.53	92.92%	91.59%
	Length	83.16	86.31%	86.14%	86.43	89.17%	88.07%
Middling	**Binary Gender**	76.64	82.47%	83.11%	77.77	84.76%	71.85%
	Gender Probability	86.37	88.60%	87.68%	81.30	86.35%	85.63%
	Brown Clusters	84.50	87.27%	86.59%	87.74	90.03%	90.10%
	POS (Spacy)	76.66	80.17%	75.38%	80.49	84.54%	79.08%
	POS (Ark)	73.86	79.06%	72.41%	80.07	85.05%	81.08%
Distant	AHST	71.71	80.17%	82.05%	55.40	68.28%	46.62%

Table 4: Scores for each individual feature on amateur (majority voting) and expert annotations.

the influence of the features listed in Table 4 for each annotator group.

Model Selection We perform a grid search over all possible feature combinations to find the best performing features. We find that the features with the highest performance are not necessarily the features with the best performance. For instance, token unigrams obtains the highest F1-score, precision, and the second highest recall on the amateur annotations, yet this feature fails to classify the minority classes.

Features We use a range of features focusing on both the textual information given in the tweets as well as extra-linguistic information including POS tags obtained using Gimpel et al. (2011) and Spacy[2].

In Table 4[3], we see that the most significant features trained on majority voted amateur annotations emphasize extra-linguistic features while the most significant features trained on expert annotations emphasize the content of the tweets.

Brown Clusters and Length We highlight the use of Brown Clusters (Brown et al., 1992) and length features (as inspired by Nobata et al. (2016)), as these are the only two features that classify the minority classes for both amateur and expert annotators. We use an in-house mapping of brown clusters, replacing unigrams with cluster identifiers.

[2]www.spacy.io

[3]Italics signify the best performing feature on expert annotations, bold signify the best performing features on amateur annotations (majority voting). These best performing features are then used for the respective "best" feature sets.

We follow Nobata et al. (2016), in their use of the length of comments in tokens, and the average length of the words in a tweet.

Author Historical Salient Terms Given the promising results obtained for sarcasm detection (Bamman and Smith, 2015), we calculate the Author Historical Salient Terms (AHST). We obtain up to 3200 tweets for each user in our data set, calculate the TF-IDF scores, and identify the top 100 terms. We then add a binary feature signifying the occurrence of each of these 100 terms.

Interestingly, this feature performs worse than any other feature. Particularly when trained on expert annotations, suggesting that hate speech may be more situational or that users engaging in hate speech, do not only, or even primarily engage in hate speech.

Gender Following the indication that gender can positively influence classification scores (Waseem and Hovy, 2016), we compute the gender of the users in our data set. To counteract the low coverage in Waseem and Hovy (2016), we use a lexicon trained on Twitter (Sap et al., 2014) to calculate the probability of gender. Using these probabilities we assign binary gender. Both the probability of a gender for a user and the binary gender are used as individual features. We find that using gender information only contributes to the classification score for amateur annotators.

Minority Class Misclassification We find that some features trained on expert and amateur annotations result in misclassification on the minority classes, including identifying no instances of the mi-

	Amateur			Expert		
Feature Set	F1	Recall	Precision	F1	Recall	Precision
Close	86.39	88.60%	87.59%	91.24	92.49%	92.67%
Middling	84.07	86.76%	85.43%	87.81	90.10%	88.53%
Distant	71.71	80.17%	82.05%	77.77	84.76%	71.85%
All	86.39	88.60%	87.59%	90.77	92.20%	92.23%
Best	83.88	86.68%	85.54%	91.19	92.49%	92.50%
Baseline	70.84	79.80%	63.69%	77.77	84.76%	71.85%

Table 5: Scores obtained for each of the feature sets.

nority classes (see Table 4). These misclassifications of the minority classes are largely due to the small number of instances in those classes. In spite of this, we do not believe that only boosting the size of the minority classes is a good approach, as we should seek to mimic reality in our data sets for hate speech detection.

Results Running our system on the Waseem and Hovy (2016) data set, we find that our best performing system does not substantially outperform on the binary classification task Waseem and Hovy (2016) ($F1_{ours}$: 70.05, $F1_{WH}$: 69.94). We find that our system performs significantly worse than Waseem and Hovy (2016) on the multi-class classification ask ($F1_{ours}$: 53.43, $F1_{WH}$: 73.89).

Interestingly, the main cause of error is false positives. This holds true using both amateur and expert annotations. We mitigate personal bias in our annotations, as multiple people have participated in the annotation process. Waseem and Hovy (2016) may suffer from personal bias, as the only the authors annotated, and only the annotations positive for hate speech were reviewed by one other person.

It is our contention that hate speech corpora should reflect real life, in that hate speech is a rare occurrence comparatively. Given that some of our features obtain high F1-scores, in spite of not classifying for the minority classes, we suggest that the unweighted F1-score may not be an appropriate metric to evaluate classification on hate speech corpora.

4 Related Work

Most related work in the field of abusive language detection has focused on detecting profanity using list-based methods to identify offensive words (Sood et al., 2012; Chen et al., 2012). These methods traditionally suffer from a poor recall and do not address hate speech. While Sood et al. (2012) incorporate edit distances to find variants of slurs, they are not able to find terms that do not occur in these lists. Nobata et al. (2016) address this, by using comprehensive lists of slurs obtained from Hatebase[4].

Waseem and Hovy (2016) and Ross et al. (2016) focus on building corpora which they annotate for containing hate speech. Our work closely resembles Waseem and Hovy (2016), as they also run classification experiments on a hate speech data set. Waseem and Hovy (2016) obtain an F1-score of 73.91 on their data set, using character n-grams and gender information.

Nobata et al. (2016) employ a wide array of features for abusive language detection, including but not limited to POS tags, the number of blacklisted words in a document, n-gram features including token and character n-grams and length features. The primary challenge this paper presents, is the need for good annotation guidelines, if one wishes to detect specific subsets of abusive language.

5 Conclusion

We find that using expert annotations can produce models that perform comparably to previous classification efforts. Our best model is on par with previous work on the Waseem and Hovy (2016) data set for the binary classification task but under-performs for the multi-class classification task.

We suggest that a weighted F1-score be applied in evaluation of classification efforts on hate speech corpora, such that misclassification on minority classes is penalized.

Our annotation and classification results expand on the claim of Ross et al. (2016) that hate speech is hard to annotate without intimate knowledge of hate speech. Furthermore, we find that considering only cases of full agreement among amateur annota-

[4]www.hatebase.org

tors can produce relatively good annotations as compared to expert annotators. This can allow for a significant decrease in the annotations burden of expert annotators by asking them to primarily consider the cases in which amateur annotators have disagreed.

Future Work We will seek to further investigate the socio-linguistic features such as gender and location. Furthermore, we will expand to more forms of hate speech. Finally, we will review the negative class in Waseem and Hovy (2016).

References

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE, September.

Adrian Chen. 2012. Inside facebook's outsourced anti-porn and gore brigade, where 'camel toes' are more offensive than 'crushed heads'. `http://gawker.com/5885714/inside-facebooks-outsourced-anti-\porn-and-gore-brigade-where-camel-\toes-are-more-offensive-than-crushed-\heads`. Last accessed on July 4th, 2016.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist eory and antiracist politics. *University of Chicago Legal Forum*, 1989(1).

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. 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*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. Bochum, Germany, September.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar, October. Association for Computational Linguistics.

Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. 2012. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS-12-06 of *AAAI Technical Report*. AAAI.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American Psychologist*, 62(4).

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, San Diego, California, June. Association for Computational Linguistics.