Automated hate speech detection in a low-resource environment

Roberts, Ethan
University of Cape Town
ethan.roberts@uct.ac.za

Abstract

The problem of hate speech on social media is a growing concern. Much work has been done to tackle online hate including work into the automated detection of hate speech. The problem of automated hate speech detection at scale, however, remains by and large unsolved. This is in part due to the difficulty of classifying short texts without contextual information, difficulties in ensuring consistent annotation quality, contextual differences in different regions and social settings, and the informal and nuanced language used on social media. Automated detection of hate speech is made all the more difficult in low-resource regions for which large existing hate speech corpora are unavailable. Here, I present a sampling framework to tackle some of these challenges. The framework uses sequential data annotation phases, each allowing for the training of a hate speech filter that further refines our ability to collect useful data in subsequent phases. This framework is implemented for two phases on Twitter data collected around discourses in South Africa, and its efficacy assessed through a cross-dataset analysis between phases, as well as an analysis to determine the classification performance of decision tree-based methods on relatively small datasets. I conclude that this framework is a viable approach for curating hate speech corpora for automated hate speech detection in a low-resource setting.

Keywords: classification, machine learning, natural language processing, hate speech detection

1 Introduction

Research into hate speech on social media is a field that has gained significant traction in recent years. The number of papers exploring the automated detection of hate speech, as well as the number of studies investigating hate speech within particular contexts has burgeoned. This growth is fuelled in part by the influx of people engaging in the online space, and the link between discourse on social media and corresponding real-world physical events. One such example is the US insurrection (Lee et al. 2022). Given the vast amount of data that needs to be analysed in order to solve this issue, machine learning can be a crucial analysis tool. However, most research into machine learning for hate speech detection is centred in the global North, and algorithmic solutions and datasets developed abroad (Poletto et al. 2021; Qian et al. 2019; Velankar et al. 2022) are not effective when deployed in South Africa because of differences in language and context specific social and political issues.

Much work has already been done using rule or lexicon-based methods to compute toxicity and sentiment (Hutto and Gilbert 2014; Loria 2018; Watanabe et al. 2018). These approaches, however, suffer from lack of understanding of contextual information, and are often subject to bias. For example, Sap et al. (2019) show that annotation of hate speech is subject to bias on the basis of vernacular, specifically, tweets written in "African-American-English" are 1.5× more likely to be flagged as offensive. Additionally, annotation quality has been shown to have a detrimental impact to classification models (Waseem 2016). Accurate annotation has also been shown to be difficult in terms of differentiating offensive language, from language that is hateful. This has been tackled in studies that examine ternary classification for of online speech (Davidson et al. 2017), though biases still arise when discerning offensive and hateful speech between target groups. Approaches to minimise systematic errors and biases have also been investigated, and are important contributions to the field. These include data augmentation (Vidgen et al. 2021) to yield improved classification performance and metrics to identify which posts are more likely to display bias between annotators (Akhtar et al. 2019), although the latter requires multiple annotators la-

belling each post.

Many studies have investigated the use of deep learning for online hate speech detection, including the use transfer learning applied to transformer models for hate speech detection (Adewumi et al. 2022; Mnassri et al. 2022; Subramaniam et al. 2022). There is, however, still little consensus regarding best approaches to automated hate speech detection, especially due to the inherently varying challenges in implementing hate speech detection in different contexts. Interestingly, Galke and Scherp (2021) have shown that Bag-of-Words (BoW) models can out perform graph-based models on text classification tasks.

An additional challenge to researchers investigating online hate speech in low-resource environments, i.e. in specific contexts where large existing datasets are unavailable, is that the cost of collecting and annotating large enough volumes of data is a significant barrier to tackling problems that are more readily engaged with in the global North. Moreover, the cost of training state-of-the-art language models is a barrier to many researchers as well (Schwartz et al. 2020).

To address these challenges, in this paper I present a framework for data collection and annotation in a low-resource setting. The framework is a means of maximising annotation resources, a costly and scarce resource, in order to generate corpora for automated hate speech detection in low-resource environments. Section 2 outlines the proposed framework and describes our data collection, Section 3 presents results on hate speech detection comparing a number of weak-learning approaches, and Section 4 discussed concluding remarks. Future work, of which there is much, is discussed in Section 5.

2 Data Collection and Methods

No single best practice exists for gathering data for online hate speech corpora. Standardising data collection strategies is however an important consideration in ensuring reproducible studies. Authors have, however, outlined broad data collection strategies for qualitative studies in this field (e.g. Gerbaudo (2016)). These include top sampling, random sampling, and zoom-in sampling. Random sampling is the only means of ensuring a statistically representative dataset and for test sets on which to determine some quantitative metrics such as fraction of hate speech, though it is difficult to implement in practice, and not suitable for compiling hate speech corpora since the fraction of hate speech is very low.

2.1 Filtering Framework

Here I outline a framework for data collection and annotation to facilitate the training of automated hate speech detection models. The framework focuses on maximising the utility of annotation resources in a way that scales from very low to high-resource data regimes. It involves sequential data annotation phases, each of which enables the development and improvement of a hate speech filter that further refines our ability to annotate useful data in subsequent phases. The hate speech filter is an algorithm that is a means of sifting through the volume of largely 'normal' online content, and preferentially selecting data that are more hateful or offensive. This means of filtering out hate speech becomes more effective with each data collection and annotation phase, resulting in a more nuanced and focused dataset on which human annotators can provide labels.

While the framework is implemented for two phases in this paper, it can in general be implemented for any number of phases. For n phases, one begins by splitting the total collected dataset into n distinct partitions. These partitions need not be the same size. This is followed by data annotation of the first phase of data, annotating for whether each post is hateful or not. Subsequent to the first annotation phase, a classifier can then be trained on the first phase of data, in order to predict the probability that an online post is hateful. This probability is a proxy for the "hatefulness" of each post. While this is done using a Random Forest Classifier trained on unigram and bigram post features in this exper-

iment, there are many possible choices for suitable text-based classifiers. The trained classifier is then used to generate predicted probabilities of hate for all posts in the next phase.

The next step is to preferentially select posts with a higher level of predicted hatefulness for annotation in order to combat the problem that data collected from social media exhibits a highly-imbalanced hate/not ratio. Using the probabilities assigned to each post, one can preferentially select data with a high probability of hate. This can be done by sampling data from this new phase, with probability of selection proportional to the probability of hate for each post. The resulting sub-sampled data from this phase can then be annotated, and the classifier retrained on all annotated data. For the i-th phase, $i \in [2, n]$, probabilities that each post is hateful are generated using a classifier trained on the previous $(i-1)$ phases.

A limitation of this sampling method for sequential phases is that systematic bias can be introduced into all subsequent phases, and on the basis of both the data and labels in the prior phases. For instance, hateful topics that are more extensively covered in the prior phases are more likely to be selected in subsequent phases, and under-represented topics are less likely to be selected. Crucially, this means that data collected in phases subsequent to the first are likely to under-represent potentially interesting forms of hate if the data in the first phase are not sufficiently representative. This has the potential to lead to blind spots in any resulting automated hate speech detection algorithms.

The choice of sampling strategy can have a huge impact on the level of systematic bias. A good choice of sampling strategy can to some extent mediate the degree to which systematic biases may arise. For example, a weak preference of selecting more hateful content will yield a lower level of bias (elaborate on the type of bias) than a strong preference. Another reasonable sampling strategy is to sample evenly from the subspace of "probability of being hate". This will result in a dataset where each item is as likely to be maximally hateful, as it is to be minimally hateful. This method of sampling, however, is also likely to yield a large degree of bias since the number of maximally hateful posts is very small. Regardless of the choice of sampling strategy, the potential benefit of covering more hateful data with limited resources needs to be weighed against the risk of introducing bias into the dataset.

In this paper, a simple approach for data selection has been taken, where the second phase of data is split into a high-hate and low-hate sample and data are sampled from each with an even probability. The threshold for differentiating high and low-hate samples is determined through a qualitative assessment of the resulting samples through trial and error. I find this ad hoc approach to be effective in practice, since the probabilities returned by the classifier are not well-calibrated and the small model size means there are clusters of posts all with identical probabilities of being hateful. The latter is illustrated in Figure 1.

This framework is implemented for data collected around specific events or flare-ups that are likely attractors of discourse and hate. While collecting data based on searches for slurs and derogatory slang is also possible, these types of posts are more easily flagged by content moderators on social media platforms, making subsequent searches for this content difficult. Whether or not this is the primary reason for the lack of this form of hate in the data we collect, this form of hate is less prevalent in both the data collected for this study as well as in the online ethnographic investigations we perform to explore discourses. In addition to collecting data around certain flare points, posts are gathered from a number of hateful users who post prolifically within these discourses and attract significant interaction, for example through retweets and replies. Identifying and following hateful users has been investigated (Horta Ribeiro et al. 2018) and is shown to be a useful approach in tackling online hate. Using the terminology discussed by Gerbaudo (2016), the proposed data collection strategy makes use of a combination of zoom-in and top sampling.

2.2 Data

The data collection for this study was undertaken by a group of researchers and annotators, and thus I refer to the collective "we" throughout this description. We collect Twitter data around a number of discourses in South Africa, specifically the protests at Brackenfell High School, and farm murders at Senekal. Both of these events gained significant traction in the mainstream media and on social media in South Africa. The dataset is comprised, by and large, of tweets written in English, though multilingual slang is commonplace. Tweets written in Afrikaans, Sesotho, Xhosa and Zulu make up a small fraction of the dataset. The exact fraction is only determined through Twitter's automated language identification, and is not accurate for low-resource or very informally written language. Annotating this dataset with a binary "hate"/"not hate" label constitutes our phase-I data. For the second phase of data collection, we consider a larger number of topics, which are outlined in Table 1. We collect tweets using a combination of the Twitter API, and Communalytic (Gruzd and Mai 2021). The advantage of using Communalytic is that it automatically collects replies and replies to replies of tweets of interest, yielding more complete conversations and therefore a complete account of the discourse of interest. Specifically, the use of Communalytic subverts to some extent the mechanism of self-selection (Tufekci 2014) people who post with a specific hashtag or using specific terms are declaring their support for a particular viewpoint or political inclination. By including replies and replies to replies, posts that are not present in the dataset as a result of this self selection mechanism are included.

These data are labelled by a group of 8 annotators, most of whom are graduate students at the University of Cape Town. This is a notable, and expensive, deviation from many studies (e.g. Burnap and Williams (2015); Davidson et al. (2017); Waseem (2016)) that outsource annotation to remote workers through platforms such as Mechanical Turk. This on its own, however, does not guarantee higher levels of annotation quality or inter-annotator agreement. The data are annotated for 2 categories, with each post analysed in part of a conversation/thread as opposed to in isolation. The first category is a binary classification of whether the post constitutes hate speech or not, and the second is an offensiveness rating on a scale of 1-10. The annotators were given instructions for the annotation process through a workshop on hate speech and annotating for hate speech, and additionally participated in weekly follow-up meetings to discuss the material. They were instructed to provide offensiveness ratings on the basis of how offensive the average reader would find the material. In terms of annotating for the "hate"/"not hate" binary category, the definitions of hate speech follow those outlined in Section 10 of the Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000 (PEPUDA). Specifically, the material needs to demonstrate a clear intention to be hurtful, cause or incite harm, and promote or propagate hatred in order to be classified as hate speech.

It is worth noting, however, that the noise on offensiveness ratings, both within annotator sets and between annotators, is too large to warrant inclusion in future studies that deal with small data volumes, and a ternary classification of hate, offensive and clean data as proposed by Davidson et al. (2017) may be more suitable.

A random forest classifier trained on the annotated phase-I data is used to generate probabilities that each post in our phase-II dataset is hateful. While more tree-based methods are investigated in Section 3, the use of boosting algorithms is omitted here in order to avoid overfitting on the small phase-I dataset. n-gram features are used for this analysis, where $n \in [1, 2]$, though it is noteworthy that term frequency-inverse document frequency (TF-IDF) features are a viable alternative. However, the resulting difference in classification performance using these two approaches is negligible, and thus only the former is reported on for simplicity. The scikit-learn implementation (Pedregosa et al. 2011) of random forest classifiers is used for this analysis, the hy-

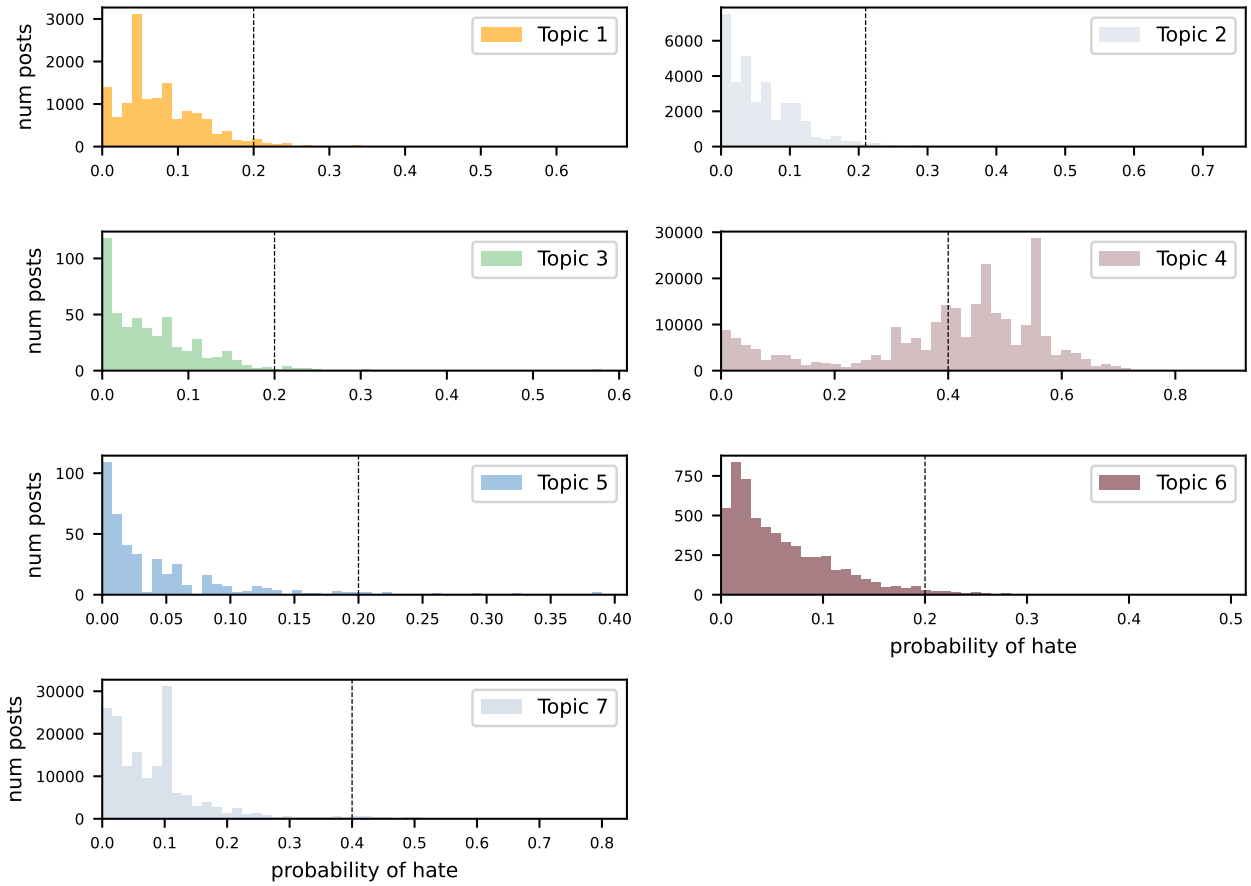

Figure 1: Histograms showing the predicted hate content for each topic outlined in Table 1 for phase-II data. The plotted probabilities are generated using a random forest classifier trained on data collected in phase-I as discussed in Section 2.2. The vertical dashed lines are the thresholds differentiating the high and low-hate samples as described in Table 1.

Table 1: Summary of topics and/or hashtags (#) that make up the phase-II dataset. The leftmost column shows the topic numbers, which correspond to the plots in Figure 1. The thresholds used to differentiate the high-hate and low-hate samples from each topic are shown in the right column. A description of how the probability thresholds are determined and used is given in Section 2.2.

	description	threshold
1	#feesmustfall	0.2
2	#phoenixmassacre	0.21
3	Jack Markovitz	0.2
4	#putsouthafricafirst	0.4
5	Brendin Horner/#farmmurders	0.2
6	#apartheidflag	0.2
7	Hateful user follow-up	0.4

perparameters for which are determined using a grid search based on classification performance averaged over a 5-fold cross-validation assessment on phase-I data alone. The resulting hyperparameters using this approach are similar to those found by simply minimising overfitting (i.e. choosing a relatively low maximum tree depth, number of estimators, and selecting a suitable value for CCP_α). It is not surprising that best performance is achieved by only (for the most part) considering minimising overfitting, given the small size of the phase-I dataset. The resulting probabilities using this approach are shown in Figure 1.

The data we collect during phase-I of this study contains 4799 posts with 9.8% flagged as hate, and phase-II contains 9115 posts with 10.8% flagged as hate. However, the subset of the phase-II data constituting the high-hate sample has 15.7% of posts flagged as hate and the low-hate sample has 5.9% flagged. This indicates that a different sampling strategy (for example, only selecting 10% of data from the low-hate sample instead of 50%) for compiling the phase-II dataset could yield a final dataset with a much larger fraction of hateful content. Additionally, one would expect this process to become more effective with the inclusion of additional data collection phases, though this investigation is left to future work.

2.3 Models and Features

Three decision tree-based classifiers are investigated in the analyses that follow: random forest classifiers, light gradient boosting machine (LightGBM) classifiers and CatBoost classifiers. In all cases, ball-park hyperparameter choices are determined using Bayesian optimisation. Details about this hyperparameter optimisation are included in Appendix A.

Similarly to the data selection described in Section 2.2, unigram and bigram features are used for the cross-dataset analysis presented in Section 3. The vocabulary size of the full dataset is 24239, and the total number of features (unigrams and bigrams) used in the analysis is 10^5, where the remaining 75761 features are the most commonly occurring bigrams in the dataset. The features used in each fold in the analyses presented in Section 3, however, have a lower resulting vocabulary, since the features are determined from the training set of that fold alone in order to avoid leakage. The full dataset used in the cross-dataset contains 13914 posts/instances, and has an average word count of 26.6 words per post.

3 Results

Reporting on the efficacy of the methodology described in Section 2 is challenging, since ideally one would compare the results using this framework with results using an equivalent set of randomly curated tweets on the same topics. The cost of annotating sufficiently large datasets to perform this analysis is however a barrier to implementing such an approach. In lieu of this, I implement two analyses, which illustrate aspects of the phased data annotation approach described in Section 2.1. The first is a cross-dataset analysis (e.g. Chen et al. (2020); Thambawita et al. (2020); Zandamela et al. (2022)) between phases-I and II, which tests aspects of the quality of the sequentially collected and curated data. The second is a typical analysis testing the classification performance of a number of deci-

sion tree-based classifiers on the full dataset (combined phases-I and II), determining which is most robustly suited to hate speech detection under relatively low volumes of data.

The phase-II dataset is processed using a random stratified sample for the cross-dataset analysis, in order to ensure a size comparable with the phase-I dataset. The sizes of the respective datasets are reported on in Section 2.2. The cross-dataset analysis consists of four sub-analyses: (1 and 2) typical analyses testing the intra-phase classification performance on phases I and II, (3) testing the classification performance by training on phase-I and testing on phase-II, and (4) testing the classification performance training on phase-II and testing on phase-I. The results of these sub-analyses are shown in Figure 2. The figure illustrates mean Matthew's Correlation Coefficient (MCC) scores, with uncertainty estimates accounting for inter-fold variation and different classification model instantiations. Sub-analyses 1 and 2 are performed using 5-fold stratified cross validation using the models described in Section 2.3. Sub-analyses 3 and 4 are performed using 5 stratified sub-samples from the phase-II dataset. Averaged results for these analyses are illustrated in Figure 2.

The results of the inter-dataset analysis, while not intended to provide a complete picture of the efficacy of a phased data collection approach, provide some degree of insight into the coherence of the sequentially collected datasets. Generating predictions on phase-II data yields worse classification performance in the above analyses. This is due to the larger number of topics covered by the phase-II dataset, and therefore a higher level of variability in the dataset.

The second analysis presented here investigates the suitability of a number tree-based classifiers for hate speech detection using the combined dataset discussed in Section 2.2. The classifiers considered are LightGBM, CatBoost and Random Forests. The performance of each classifier is evaluated using 5-fold cross validation over the combined dataset. In each fold, however, 10% of the data are held out for validation in order to determine an appropriate classification threshold. This is an important step, since the trained classifiers are highly uncalibrated without the use of calibration methods such as Platt scaling or isotonic regression. The results of this analysis are summarised in Table 2.

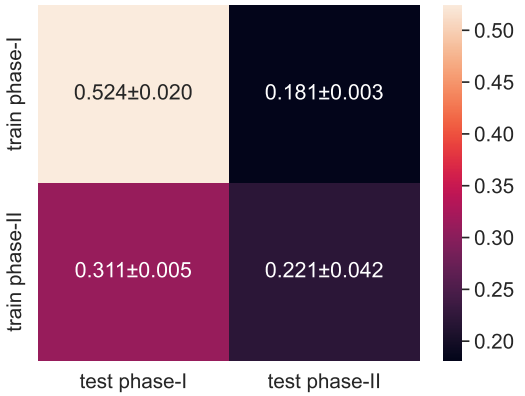

Figure 2: Cross-phase MCC scores illustrating classification performance for the cross-dataset analysis described in Section 3. The scores shown are mean scores of each fold in a classification analysis using 5-fold cross validation, and the uncertainty estimates are the standard deviations of the scores. A score of 1 shows perfect classification, and a score of 0 corresponds to random guessing.

Table 2: Classification results illustrating the performance of a number of tree-based classifiers for hate speech detection. The metrics considered are Matthew's Correlation Coefficient (MCC) and accuracy. When evaluating classification models with imbalanced datasets, MCC is a more informative metric than accuracy as it is insensitive to class imbalance. MCC scores of 1 correspond to perfect classification, 0 to random guessing, and -1 to total incorrect classification.

Classifier	MCC	Accuracy (%)
LightGBM	0.320 ± 0.006	87.29 ± 0.11
CatBoost	0.318 ± 0.020	87.22 ± 0.38
RandomForest	0.232 ± 0.025	85.53 ± 0.85

LightGBM shows superior classification performance under both metrics considered, though the performance of CatBoost is within the margin of error. In addition to this, I find the computational cost of LightGBM to be significantly lower than that of CatBoost and Random Forests. A summary of the computational costs for each algorithm is shown in Table 3. This difference in computational performance is due primarily to the different tree-growth strategies used by the respective algorithms. Specifically, LightGBM uses a leaf-wise as opposed to a level-wise splitting strategy, which allows for the more efficient development of accurate classification trees. This is especially notable when the number of features in the data are very large. While this tree-growth strategy can increase the chance of overfitting, this can be somewhat mitigated through a limitation of maximum tree depth and use of regularisation.

Table 3: Train and test times for the classifiers considered in this analysis. The reported times are for CPU algorithm implementations on a single fold (of a 5-fold validation) on the full dataset. This analysis was done on a 2.9GHz Intel Quad-Core i7 processor.

Classifier	Train time (s)	Test time (s)
LightGBM	20	2
CatBoost	554	60
Random Forests	590	2

In addition to the classification of hate speech, I also analyse the predictive performance using the offensiveness scores described in Section 2.2. However, as described in Section 2.2, the noise on labels and sparse textual features make generating predictions of offensiveness difficult. This is quantified through a regression analysis using a random forest regressor trained on the same textual features as in the above analysis. The results culminate in a Mean Squared Error (MSE) of 4.159 ± 0.124 and an R^2 of 0.094 ± 0.012. Considering the poor performance using the offensiveness scores, further analysis using these scores is not included in this work.

4 Concluding remarks

Accurate automated detection of online hate remains a highly complex task, even when relevant large training datasets are available. This is in part due to the fact that short posts on social media are highly subject to context and nuance, making their classification challenging even to human annotators.

Overall, the framework put forward in this work shows a systematic approach to the problem of automated hate speech detection, and provides a method of generating an annotated hate speech corpus in a low resource environment. The performance achieved in the analysis presented in Section 3 illustrates the efficacy of this framework using only simple text-based features. This shows that this framework is a viable approach for tackling online hate in specific geographic or political contexts, where existing large datasets are unavailable. The cross-dataset analysis I present illustrates coherence between the different data collection phases in both directions, indicating that this approach is viable for maximising annotation resources in any hate speech detection study.

While the data curated for the second phase of the study does not have a markedly higher hate fraction than the initially collected data (10.8% vs 9.8%), the annotators report that the second phase of data was far more hateful on average. This qualitative assessment warrants further investigation, especially since disentangling hateful and simply offensive language is an important consideration in the curation of online hate speech corpora (Davidson et al. 2017; Watanabe et al. 2018). Additionally, investigation into the high-hate sample in the phase-II data shows that the hate fraction could be increased to as high as 15.7% with a different selection strategy.

Furthermore, I conclude that LightGBM offers both the best and most practical approach to decision-tree based hate speech classification, and is the most suitable candidate, of those considered here, for comparison to more sophisticated methods in a low-resource setting.

5 Future work

The problem of automated hate speech detection on social media is one that is difficult to tackle, owing in part to the diversity of the underlying data in different contexts and the subjectivity in human annotation. Though much research has already been done on this topic, there remains substantial scope for further work.

Some of these extensions are outlined below:

1. Additional data collection and annotation phases would allow for a more quantitative assessment of whether this framework yields improved automated hate speech detection compared to standard data collection techniques. A larger dataset is also imperative in the training of more sophisticated analysis tools, such as Large Language Models (LLMs).
2. Transfer learning/fine-tuning of LLMs for hate speech detection has already received a fair amount of attention in the literature (Adewumi et al. 2022; Mnassri et al. 2022; Subramaniam et al. 2022). This approach leverages a large amount of pre-training to produce hate speech detection models using a comparatively small volume of hate speech data. Investigating the threshold for hate speech data volume at which this approach outperforms traditional keyword-based or ML-based methods is a topic of interest to many researchers tackling online hate in low-resource regions.
3. Topic modelling and content understanding are important considerations in analysing online discourse and hate speech. The difficulties of implementing these tools on noisy and complex social media data mean that their applications remain largely unexplored. Incorporating sophisticated ML tools with more comprehensive annotation schemes, has the potential to both deepen analyses, as well as offer potential interventions to online harm (e.g. see Masud et al. (2022)).
4. Investigations of inter-annotator reliability are important both in understanding biases inherent in the resulting classification models, and in developing more robust annotation practices. Collecting a larger volume of data where multiple annotators label each post would allow for analyses of inter-annotator reliability.

These extensions are left to future work.

Acknowledgements

I sincerely thank the reviewers for their comments, which yielded valuable additions to this paper. Though the work presented in this paper is that of a single author, it is part of a larger project involving a group of team members without whom, this work would not be possible. I thank Adam Mendelsohn, Gavaza Maluleke and Thierry Rousset for their oversight of the entire project, for their management of data collection and annotation, and for ongoing discussions in tackling online hate. The work that they have done on the exploration of hate on social media in South Africa can be found in this report Rousset et al. (2022). I also thank the team of 8 annotators who painstakingly sifted through mountains of social media data, resulting in the datasets that are the backbone of this work.

Appendix A: Hyperparameter optimisation

Details of the hyperparameter optimisation described briefly in Section 2.3 are shown in the tables below. Bayesian optimisation is used to fit hyperparameters for the algorithms considered in this work. Tables 4, 5 and 6 show which hyperparameters are optimised for the analysis with Random Forests, LightGBM and CatBoost respectively.

Table 4: Hyperparameter optimisation specifications for random forests using Bayesian optimisation. Tuned parameters are shown as a range in the value column and fixed parameters are specified as a single number. Prior distributions on discrete parameters are uniform by default, and therefore not specified here.

Random Forests		
Hyperparameter	Value	Prior
n_estimators	200-2000	
max_depth	3-10	
min_samples_split	1-2	
min_samples_leaf	1-2	
ccp_alpha	0.0-1.0	uniform
min_impurity_decrease	0.0	

Table 5: Hyperparameter optimisation specifications for LightGBM using Bayesian optimisation. Tuned parameters are shown as a range in the value column and fixed parameters are specified as a single number. Prior distributions on discrete parameters are uniform by default, and therefore not specified here.

LightGBM		
Hyperparameter	Value	Prior
n_estimators	200-2000	
learning_rate	0.001-0.1	log-uniform
max_depth	3-10	
num_leaves	80-800	
min_split_gain	0.0-0.3	uniform
reg_alpha	0-1.5	uniform

Further improvements to the choice of hyperparameters can however be made through following commonly prescribed procedures for reducing overfitting, such as setting a regularisation term and reducing the maximum tree depth. The importance of mitigating overfitting and the difficulty in minimising overfitting with standard hyperparameter optimisation methods, such as Bayesian optimisation or grid-search, are notable considerations for automated hate speech detection with decision tree-based classifiers.

Table 6: Hyperparameter optimisation specifications for CatBoost using Bayesian optimisation. Tuned parameters are shown as a range in the value column and fixed parameters are specified as a single number. Prior distributions on discrete parameters are uniform by default, and therefore not specified here.

CatBoost		
Hyperparameter	Value	Prior
n_estimators	200-2000	
max_depth	3-10	
learning_rate	0.001-0.1	log-uniform
l2_leaf_reg	1.0-3.0	uniform
bootstrap_type	Bayesian	
bagging_temperature	0.5	
random_strength	0.5	

References

Adewumi, T., Sabah Sabry, S., Abid, N., Liwicki, F. and Liwicki, M. (2022), 'T5 for Hate Speech, Augmented Data and Ensemble', *arXiv e-prints* p. arXiv:2210.05480.

Akhtar, S., Basile, V. and Patti, V. (2019), A new measure of polarization in the annotation of hate speech, *in* 'AI*IA'.

Burnap, P. and Williams, M. L. (2015), 'Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making', *Policy & Internet* **7**(2), 223–242.
URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.85`

Chen, Y., Liu, P., Zhong, M., Dou, Z.-Y., Wang, D.,

Qiu, X. and Huang, X. (2020), CDEvalSumm: An empirical study of cross-dataset evaluation for neural summarization systems, *in* 'Findings of the Association for Computational Linguistics: EMNLP 2020', Association for Computational Linguistics, Online, pp. "3679–3691".
URL: https://aclanthology.org/2020.findings-emnlp.329

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017), 'Automated Hate Speech Detection and the Problem of Offensive Language', *arXiv e-prints* p. arXiv:1703.04009.

Galke, L. and Scherp, A. (2021), 'Bag-of-Words vs. Graph vs. Sequence in Text Classification: Questioning the Necessity of Text-Graphs and the Surprising Strength of a Wide MLP', *arXiv e-prints* p. arXiv:2109.03777.

Gerbaudo, P. (2016), 'From data analytics to data hermeneutics. online political discussions, digital methods and the continuing relevance of interpretive approaches', *Digital Culture and Society* **2**.

Gruzd, A. and Mai, P. (2021), 'Communalytic: A Research Tool For Studying Online Communities and Online Discourse. Available at https://Communalytic.com'.

Horta Ribeiro, M., Calais, P. H., Santos, Y. A., Almeida, V. A. F. and Meira, Wagner, J. (2018), 'Characterizing and Detecting Hateful Users on Twitter', *arXiv e-prints* p. arXiv:1803.08977.

Hutto, C. and Gilbert, E. (2014), 'Vader: A parsimonious rule-based model for sentiment analysis of social media text', *Proceedings of the International AAAI Conference on Web and Social Media* **8**(1), 216–225.
URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Lee, C. S., Merizalde, J., Colautti, J. D., An, J. and Kwak, H. (2022), 'Storm the capitol: Linking offline political speech and online Twitter extra-representational participation on QAnon and the January 6 insurrection', *Frontiers in Sociology* **7**.
URL: https://www.frontiersin.org/articles/10.3389/fsoc.2022.876070

Loria, S. (2018), 'TextBlob Documentation', *Release 0.15* **2**.

Masud, S., Bedi, M., Aflah Khan, M., Shad Akhtar, M. and Chakraborty, T. (2022), 'Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization', *In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)* pp. 3524–3534.

Mnassri, K., Rajapaksha, P., Farahbakhsh, R. and Crespi, N. (2022), 'BERT-based Ensemble Approaches for Hate Speech Detection', *arXiv e-prints* p. arXiv:2209.06505.

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

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C. and Patti, V. (2021), 'Resources and benchmark corpora for hate speech detection: a systematic review', *Language Resources and Evaluation* **55**(2), 477–523.
URL: https://doi.org/10.1007/s10579-020-09502-8

Qian, J., Bethke, A., Liu, Y., Belding, E. and Wang, W. Y. (2019), A benchmark dataset for learning to intervene in online hate speech, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. "4755–4764".
URL: https://aclanthology.org/D19-1482

Rousset, T., Maluleke, G. and Mendelsohn, A.

(2022), 'The dynamics of racism, antisemitism and xenophobia on social media in South Africa'.

Sap, M., Card, D., Gabriel, S., Choi, Y. and Smith, N. A. (2019), The risk of racial bias in hate speech detection, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. "1668–1678".
URL: https://aclanthology.org/P19-1163

Schwartz, R., Dodge, J., Smith, N. and Etzioni, O. (2020), 'Green AI', *Communications of the ACM* **63**, 54–63.

Subramaniam, A., Mehra, A. and Kundu, S. (2022), 'Exploring Hate Speech Detection with HateXplain and BERT', *arXiv e-prints* p. arXiv:2208.04489.

Thambawita, V., Jha, D., Lewi Hammer, H., Johansen, H. D., Johansen, D., Halvorsen, P. and Riegler, M. A. (2020), 'An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning applied to Gastrointestinal Tract Abnormality Classification', *ACM Trans. Comput. Healthcare* **1**.

Tufekci, Z. (2014), 'Big questions for social media big data: Representativeness, validity and other methodological pitfalls', *Proceedings of the International AAAI Conference on Web and Social Media* **8**(1), 505–514.
URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14517

Velankar, A., Patil, H. and Joshi, R. (2022), 'A Review of Challenges in Machine Learning based Automated Hate Speech Detection', *arXiv e-prints* p. arXiv:2209.05294.

Vidgen, B., Thrush, T., Waseem, Z. and Kiela, D. (2021), Learning from the worst: Dynamically generated datasets to improve online hate detection, *in* 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Online, pp. "1667–1682".
URL: https://aclanthology.org/2021.acl-long.132

Waseem, Z. (2016), Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter, *in* 'NLP+CSS@EMNLP'.

Watanabe, H., Bouazizi, M. and Ohtsuki, T. (2018), 'Hate speech on twitter a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection', *IEEE Access* **PP**, 1–1.

Zandamela, F., Ratshidaho, T., Nicolls, F. and Stoltz, G. (2022), 'Cross-dataset performance evaluation of deep learning distracted driver detection algorithms', *MATEC Web of Conferences* **370**.