Ahmed et al. *EPJ Data Science* (2022) 11:8
https://doi.org/10.1140/epjds/s13688-022-00319-9

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning

Zo Ahmed[1], Bertie Vidgen[1,2] and Scott A. Hale[1,2,3*]

*Correspondence:
scott.hale@oii.ox.ac.uk
[1]Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK
[2]Alan Turing Institute, 2QR, 96 Euston Road, London, NW1 2DB, UK
Full list of author information is available at the end of the article

Abstract

Online hate is a growing concern on many social media platforms, making them unwelcoming and unsafe. To combat this, technology companies are increasingly developing techniques to automatically identify and sanction hateful users. However, accurate detection of such users remains a challenge due to the contextual nature of speech, whose meaning depends on the social setting in which it is used. This contextual nature of speech has also led to minoritized users, especially African–Americans, to be unfairly detected as 'hateful' by the very algorithms designed to protect them. To resolve this problem of inaccurate and unfair hate detection, research has focused on developing machine learning (ML) systems that better understand textual context. Incorporating social networks of hateful users has not received as much attention, despite social science research suggesting it provides rich contextual information. We present a system for more accurately and fairly detecting hateful users by incorporating social network information through geometric deep learning. Geometric deep learning is a ML technique that dynamically learns information-rich network representations. We make two main contributions: first, we demonstrate that adding network information with geometric deep learning produces a more accurate classifier compared with other techniques that either exclude network information entirely or incorporate it through manual feature engineering. Our best performing model achieves an AUC score of 90.8% on a previously released hateful user dataset. Second, we show that such information also leads to fairer outcomes: using the 'predictive equality' fairness criteria, we compare the false positive rates of our geometric learning algorithm to other ML techniques and find that our best-performing classifier has no false positives among a subset of African–American users. A neural network without network information has the largest number of false positives at 26, while a neural network incorporating manual network features has 13 false positives among African–American users. The system we present highlights the importance of effectively incorporating social network features in automated hateful user detection, raising new opportunities to improve how online hate is tackled.

Keywords: Hateful users; Geometric deep learning; Algorithmic bias

© The Author(s) 2022. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Springer

1 Introduction

The massive expansion in social media over the last two decades has brought unprecedented connectivity and communication. Unfortunately, such communication is sometimes characterized by harm and abuse, such as hate speech. This hate not only negatively impacts users of online platforms and their communities [54], it can also stir up social tensions and affect the reputation of the platforms who host them [56]. Online hate can have implications in the offline world: it has been linked to youth suicides, mass shootings, violent hate crimes, and extremist recruitment [26].

Yet, the problems of accuracy and bias faced by hateful content detection also plague hateful user detection. Hate, by its very nature, is context-dependent: it depends, at a minimum, on the intention of the potential perpetrator and the reception of the alleged target [38]. Language, moreover, can be 'polysemous', with different meanings in different contexts, increasing the vulnerability of machine learning (ML) systems to bias if they lack adequate context-awareness [48]. Words, such as 'queer' or 'nigga', can be used hatefully in some contexts. In others, however, they may be used to describe one's own identity, i.e. self-referentially, or serve as an innocuous part of a group's dialect [3]. While more contextual Natural Language Processing (NLP) methods like Bidirectional Encoder Representations from Transformers (BERT) have helped alleviate some problems, automated hate detection still suffers from being inaccurate and biased towards minoritized groups, especially African–Americans [45, 55]. Despite the wealth of social science research on how hateful users online are embedded in larger network structures, the incorporation of network information in hate detection has been extremely limited. Hateful users have been shown to operate in close-knit networks online [13, 53], while such homophily also exists on racial and ethnic lines [35]. By enabling models to incorporate more informative network-level signals, more accurate hate detection may be possible. At the same time, including network-data may also help ML models better situate text in its social context, leading to fairer outcomes for minority group users who may share similar linguistic characteristics with hateful users.

We examine how different approaches to incorporating network information affect both the accuracy and fairness of ML models to detect hateful users. We use a partially annotated retweet network dataset containing up to 200 tweets per user for 100,000 unique users, provided by Ribeiro et al. [53], and use geometric deep learning to incorporate network-data in a learned end-to-end manner. We compare this technique to other algorithms that either ignore network data or use manually engineered network-level features. We find that our geometric deep learning based systems result in fairer and more accurate hateful user classifications. Our work suggests that network information, when incorporated effectively, can provide the additional context required for fairer *and* more accurate automated hate detection.

In Sect. 2 we explain the benefits of incorporating social network features by reviewing research on the social dynamics of hateful behaviour online. We also appraise existing automated approaches for hate detection, highlighting their context-blindness and the challenges they face in producing fair and accurate outcomes. We then review the concept of fairness in machine learning, and the particular issues of bias for automated hate detection. In Sect. 3 and Sect. 4 we outline the data and methods, describing the labelled dataset from Ribeiro et al. [53] and methods for detecting hateful users and identifying African–Americans. In Sect. 5 we present and compare the results of various classifica-

tion models for their overall performance, as well as their performance on the subset of African–American users. We also conduct a detailed error analysis to diagnose the best-performing model's weaknesses. In Sect. 7 we discuss our findings and the limitations of the research.

2 Related work

Hate is a social phenomenon whose characteristics we highlight in Sect. 2.1, with emphasis on how its complexities relate to the difficulties faced in accurately and fairly detecting online hate using ML systems. Existing research on automated hate detection systems is next described in greater depth in Sect. 2.2, with progress and persistent challenges in the accuracy of such systems underlined. Finally, background is provided into the fairness challenges faced by such systems in Sect. 2.3, with relevant research on the nature, implications, and possible solutions of these challenges highlighted.

2.1 The social dynamics of online hate

Online hate is a contested and context-dependent concept, and despite growing concerns about its harmful effects, platforms, governments and researchers have been unable to reach a common definition [60]. Academic work displays a fragmented understanding of online hate [37], although most definitions share three common elements: content, intent, and harm [38]. Content relates to the textual, visual or other modes used to proliferate hate. Intent and harm are more difficult to discern, and relate to the thoughts and experiences of the perpetrator and victim [47, 49]. In line with much previous work, we define online hate in line with Davidson et al. [12] as 'language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group.'

Social psychological work on prejudice and abuse proposes that the expression of hate is strongly affected by intergroup dynamics, such as contact and conflict between competing social groups [9]. Following Crandall and Eshelman's Justification–Suppression model of prejudice, Day [13] finds racism on Twitter flourishes when the offender is motivated by a group of like-minded racist peers, when prejudice against targets can be socially justified, and when capable guardians are not present to suppress their activities. It has been suggested that community ties also operate within far-right online communities, motivating and deepening individuals' commitment to extremist politics [5, 15].

Hateful users are often homophilous and organize themselves into densely connected and clustered networks where hateful ideas and content proliferate quickly [26, 53]. Johnson et al. [26] argue that content removal proves largely ineffective in such settings: the sources of the content—hate-spewing, highly resilient clusters of users who are committed to their goals—live on, growing larger with time due to consolidation of smaller groups or attraction of new recruits. Further, the growing user bases of platforms that have lenient content moderation policies and/or were created with the stated goal of protecting freedom of expression, such as Gab, Parler, and Bitchute, reflects the need to understand hateful networks and communities, rather than hateful individuals in isolation. Other forms of homophily also exist online, which could be used to understand the spread of online hate. Wimmer and Lewis [66] find that African–American individuals, for example, are more likely to associate themselves with other African–Americans on social media platforms, creating a "high degree of racial homogeneity" in social networks [66].

2.2 Automated methods of online hate detection

Hateful content detection has been extensively studied, with numerous papers presenting new datasets, models and architectures for hateful content detection, including several subtasks, such as multilingual, multimodal and multi-target hate detection [29, 61, 68]. Notably, incorporating sentence-level context has been shown to improve hateful *content* detection, with several papers deploying transformer-based models such as Bidirectional Encoder Representation from Transformers (BERT) to better distinguish between hateful and non-hateful content, even when they have lexical similarities [37, 57]. Indeed, as Mullah and Zainon [46] write in their comprehensive review of ML methods for automated hate speech detection, deep learning techniques leveraging such language models can considerably improve how context-dependent hate speech is detected. Liu et al. [36], for example, leverage transfer learning by fine-tuning a pre-trained BERT model to produce the most accurate classifier for offensive language in the SemEval 2019 Task 6 competition. The model, however, was less effective in identifying the type of the offence or the target. Chiril et al. [8], meanwhile, focus more on the target-oriented and topical focus of hate speech using a multi-target perspective, building classifiers to simultaneously categorize both the hateful nature of a text as well as its topical focus. They find that leveraging BERT under such a multi-task approach outperforms single-task models for detecting both the hatefulness of a text and its focus. Mishra et al. [43] also apply multi-task learning, back-translation, and multilingual training to the problem of hateful content detection.

However, even with such advanced contextual language models, several problems emerge. Liu et al. [36] cite the imbalanced nature of the target (hateful speech is much rarer than non-hateful speech) as well as the morphological nature of language as challenges to achieving higher performance. Meanwhile, data imbalance in the form of bias towards certain topics and targets can also hamper performance while adding the possibility of unfair classification [8]. Moreover, despite improving performance overall, such contextual language models are often still unable to distinguish complex forms of hate, such as understanding hateful uses of polysemous words related to minority groups [45]. Part of the challenge is that the meaning of such terms depends on the identity of the speaker and how they are used (e.g., a black person using the term "nigga" is fundamentally different to a white person using it). With short text statements (e.g., tweets) it can be difficult to capture and represent these different uses (particularly if the labelled training dataset is small), even with transformer-based models.

Creative ways have more recently been proposed to deal with such polysemy by better leveraging the social characteristics of online hate. Alorainy et al. [1] highlight the risk of false negatives when detecting subtle forms of hate, as traditional embedding techniques still depend heavily on the occurrence of certain words or phrases in classifying hate. Instead, they leverage social science findings around how hate is conveyed, specifically on the phenomenon of 'othering' whereby "hateful language is used to express divisive opinions between the in-group ('us') and the out-group ('them')". By augmenting traditional embedding algorithms with features specifically related to 'othering', such as pronoun patterns and verb–pronoun combinations, they achieve a significant improvement in accuracy [1].

Previous research also includes classifying online hate by either using traits at the user-level or directly classifying hateful users. Given the vast body of research showing how online hate is often propagated by a small, highly-connected group of users (see [26]), there has been increased attention on leveraging user-level information for the task. At

the 2021 Conference and Labs of the Evaluation Forum (CLEF), for instance, 66 academic teams participated in the task of directly detecting whether a Twitter user is likely to spread hate or not [51]. While the primary focus of the participants still seemed to have been the textual elements in the tweets of these users, some researchers also leveraged innovative strategies to extract and use richer information directly at the user level in the classification task. Irani et al. [24], for example, learned user-level embeddings with unsupervised and semi-supervised techniques, before combining them with textual features to more accurately classify hateful users in the provided dataset. However, few, if any, of the participants in the task incorporated network-level information when learning user-level embeddings, despite the inherent networked nature of social media [51].

Though still focused primarily on automated hateful *content* classification, a small but growing body of research leverages users' networks, using simple features such as in- and out-degree or shallow-embedding methods such as node2vec [7, 41]. Fehn Unsvåg and Gambäck [17] create a classifier in which textual features are augmented with user-specific ones, such as gender, activity and network connectivity (number of followers and friends). They use Logistic Regression (LR), reporting a 3 point increase in the F1 score with the addition of user-specific features. Mishra et al. [41] use a variant of the word2vec skipgram model by Mikolov et al. [40] to create network (node2vec) embeddings of users based on their network positions and those of their neighbours. They concatenate the generated user network vectors with character n-gram count vectors, and use a LR classifier, achieving a 4 point increase in the F1 score compared to the same model without the network embeddings.

One drawback of node2vec embeddings, however, is that they can only incorporate the structural and community information of a user's network; the textual content of tweets written by other users in this network is not incorporated. This may, as Mishra et al. [42] demonstrate, lead to misclassification of a normal user's content because of their hateful network neighbourhood through 'guilt-by-association'. This is a problem for use of a node2vec model in the real world, where such mistakes would spark considerable backlash and opposition. Most importantly, from a practical perspective, shallow embeddings such as node2vec may be unfeasible as the embeddings need to be re-generated for the entire graph every time a new user enters the network [21]. The large and constantly changing user bases of social media platforms render such methods computationally impractical.

Moving from detecting hateful content to detecting hateful users, meanwhile, is not straightforward due to challenges in sampling [12, 53, 62] and in accessing network and other user attributes [25]. Geometric deep learning offers a very promising avenue for hateful user detection, and has been used to leverage network features in other classification problems. Ribeiro et al. [53] were the first to use geometric deep learning for hateful user detection, using GraphSAGE, a semi-supervised learning method for graph structures such as networks. It incorporates a user's neighbourhood network structure as well as the features of her neighbours to learn node embeddings for unseen data [21]. Ribeiro et al. [53] achieve an overall unbalanced accuracy of 84.77% and F1-score of 54%, a substantial improvement over their baseline model. However, one limitation of their work is the relative simplicity of the baseline, which is a decision tree that lacks any network information. This is not a like-for-like comparison as the geometric deep learning model has access to additional features.

Further, they do not consider the implications of geometric deep learning beyond just accuracy: network-level features may, theoretically at least, be less biased towards the 'guilt-by-association' sub-problem because of how they exploit homophily in social networks. This is a key concern; if this fatal limitation of incorporating network features cannot be overcome then the method cannot be reliably used in real-world content moderation.

2.3 Fairness in automated hate detection

Attention in online hate research has increasingly shifted from solely considering *performance* to also evaluating models based on their *fairness*. This is particularly important for online hate detection as biased models could risk perpetuating the very problems (of social discrimination and unfairness) that they are designed to challenge.

Fairness in online hate detection has been addressed in several prior studies. Thus far, fairness has largely been explored in relation to *content*, although it is likely that it also affects hateful *user* detection algorithms, presenting the potential for new sources of harm. To our knowledge fairness in hateful user detection has not been systematically studied in previous work. Chung [10] show that Google's Perspective toxicity classifier disparately impacts African–American users by disproportionately misclassifying texts in African–American Vernacular English (AAVE) as toxic. Exclusively learning from textual elements, the model identifies relationships between words such as 'black', 'dope', or 'ass' and toxicity as such words appear frequently in hateful posts. Given the frequent use of such words in AAVE, African–American users bear the negative implications of such erroneous learning.

Davidson et al. [11] use machine learning to identify AAVE tweets and find substantial and statistically significant racial disparity in the five datasets they investigate. AAVE tweets are, for models trained on some datasets, more than twice as likely to be associated with hate compared to Standard American English (SAE) tweets. They investigate how tweets containing 'nigga', a term re-appropriated by the African–American community but still frequently used to perpetrate hate, are classified. They find that a tweet's dialect, independent of whether it uses 'nigga', strongly influences how it's classified: AAVE tweets are more likely to be classed as hate-related than SAE tweets, even where both contain the term 'nigga' [11].

Davidson et al. [11] argue that annotation errors in the training data, where black-aligned language may be more likely to be labelled as hateful, may explain some of these differences. Sap et al. [55] have confirmed this possibility, exploring two key hate speech datasets and finding that AAVE tweets are up to two times more likely to be labelled as offensive by human annotators compared to non-AAVE tweets. Models trained on such datasets learn and propagate these annotation biases, amplifying disparate impact. Even BERT-based classifiers can confuse the meaning of words such as 'women', 'nigga' or 'queer', falsely flagging non-hateful content that uses such words innocuously or self-referentially [45].

Fairness, alongside other concepts such as bias and discrimination, is contested in ML literature. Mehrabi et al. [39] provide a comprehensive review of work on 'Fair ML', and define fairness as "the absence of any prejudice or favouritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making". However, they acknowledge that fairness can have varying notions both within and across disciplines, depending on how it is operationalized. Barocas and Hardt [2] have categorized

notions of fairness into two groups: those centred around 'disparate treatment' and those around 'disparate impact'. Disparate treatment characterizes decision-making if a subject's sensitive attributes, such as race, partly or fully influence the outcome. In contrast, disparate impact occurs where certain outcomes are disproportionately skewed towards certain sensitive groups [2]. Elsewhere, Kleinberg et al. [30] show the constraints imposed by different definitions of fairness on ML systems are incompatible: in most practical applications ML models cannot optimize for a particular definition of fairness without also causing unfair outcomes when considered from a different definition. Thus, it is important to establish the notion of fairness that is being prioritized before models are optimized.

Trying to maximize algorithmic fairness can lead to a 'fairness–accuracy trade-off', where "satisfying the supplied fairness constraints is achieved only at the expense of accuracy" [65]. However, Wick et al. [65] have also shown that theoretically and practically, such a trade-off may be overcome in cases where systemic differences in the outcome variable (e.g., online hate) are not caused by the sensitive attribute (i.e., race). This is a plausible assumption in the current setting: nothing suggests that people of a particular race are intrinsically more hateful than those of another, especially after accounting for other factors. As such, it could be possible for both fairness and accuracy to be optimized, given an effective model.

Fairness can be operationalized in different ways, each of which could be justified in certain situations [39]. One notion of fairness is demographic parity, which requires that a model's predictions are the same across different sensitive groups [22]. While applicable in certain contexts, this may be suboptimal for hate detection tasks. The social reality of online hate means that certain groups are more likely to perpetrate it, such as white supremacists compared with other users (of any race) [11]. However, because demographic parity does not incorporate accuracy of classifications (i.e., whether predicted outcomes match actual labels), it could theoretically be 'achieved' with many misclassifications.

The notion of 'predictive equality' assesses the false positive error rate balance, and can also be used to evaluate fairness [58]. Predictive equality is very appropriate to studying online hate, which has a history of disproportionate false positives among minority groups. There is also a magnified harm of such errors compared to false negatives, given that it penalizes the very groups that hate classifiers are meant to protect [14].

A previously unstudied way of effectively and efficiently addressing fairness in online hate detection could be through incorporating user-specific information by leveraging their social networks. Given homophily in such networks, ML models incorporating network information might learn user characteristics without these being manually specified. Indeed, new approaches to 'Fairness without Demographics' in ML have sought to exploit correlations between observed features and unobserved sensitive group attributes to boost fairness [32]. When social networks are effectively incorporated into ML models, this homogeneity can be exploited to infer contextual information around an individual's identity, without directly extracting and feeding such features into the model. This may not only lead to more scalable models for online hate detection, but also fairer and more accurate ones. For instance, the text of a user can be contextualized against their identity, helping resolve ambiguities stemming from the analysis of text alone, such as with polysemous words like 'nigga'. This approach also has the clear advantage of minimizing the

amount of sensitive data that needs to be harvested, assessed, and stored, minimizing the risk of privacy-invasions.

3 Data

The data used in this paper was collected and labelled by Ribeiro et al. [53]. It consists of a Twitter retweet network, with a total of 100,386 users (nodes), and 2,286,592 edges (including self-loops) between them. We removed all self-loops from the dataset. The network is a directed graph $G = (V, E)$. Each node $u \in V$ represents a user and a directed edge between two users $(u_i, u_j) \in E$ denotes that u_i has retweeted u_j. As such, the edge directions can serve as a proxy for the direction of influence flow in the graph [53]. To construct a sampled graph of all Twitter users Ribeiro et al. [53] used a Directed Unbiased Random Walk (DURW) algorithm. DURW avoids oversampling high-degree nodes, producing a more representative sample [52]. The starting node was set to a Twitter account geo-located in San Jose, California.

Ribeiro et al. [53] selected 4,972 users for manual annotation (approximately 5%). The relative sparsity of hateful online content has led many researchers to use a lexicon-based approach to ensure that the sample contains more hateful content than randomly found online (e.g., [6, 12]). However, such approaches would bias sampling towards only those users who use the words in the specified lexicon, missing other types of hateful users. Indeed, notably, the entire dataset of 1,972 racist tweets collected by Waseem and Hovy [64] came from only 9 users. To mitigate this, Ribeiro et al. [53] used a graph diffusion process based on a lexicon of words mainly used to express hate, taken from databases provided by Hatebase.org and the ADL. The graph diffusion process assigns probabilities to each user based on their tweets and those of their nearest neighbour. It ensures a more balanced sample between hateful and innocuously used words in the lexicon as well as inclusion of users whose neighbours are hateful making it more likely that coded or non-standard hate is also included.

Ribeiro et al. [53] use crowdsourcing to manually annotate 4,972 of the sampled users as hateful or normal based on their sampled tweets and in reference to Twitter's guidelines on hateful conduct. Each user was annotated by at least three independent annotations. If there was any disagreement then up to five annotations were collected [53]. 544 hateful (10.9%) and 4,438 normal (89.1%) users were identified.

For each user, Ribeiro et al. [53] extracted textual, user-level, and network-level features. The text of users' tweets are represented as 300-dimensional GloVe vector embeddings, which were provided with the dataset. Other textual features were also derived, including sentiment and subjectivity scores. User-level features relate to their Twitter activity and profile, and include the number of followers, followees, statuses, and favourites normalized by the time since account creation. Network-level features include the in-degree, out-degree, eigenvector, and betweenness centrality of each user, and aggregated user-level and textual attributes of their immediate 1-hop neighbourhoods. Ribeiro et al. [53] find large and statistically significant differences existed between normal and hateful users on such measures, with hateful users being identified as more active, central, and with newer accounts, likely because of previous account suspensions. In addition, we used the original tweets for a partial set of the users, which were provided to us by Ribeiro et al. [53].[1] We

[1] We thank the original authors for providing us with this information.

used these tweets only to examine classification errors and for demography detection. We did not use them for modelling as they were not available for all users.

4 Experimental setup and methodology

The research methods used in this paper can be divided into two main categories: methods for inferring users' demographics and machine learning techniques for the automatic detection of hateful users. Following a summary of the experimental setup, a detailed explanation of each method is provided.

4.1 Experimental setup

After identifying a subset of African–American users from the dataset, as detailed in 4.2, 5-fold cross-validation is used to test both the predictive power and fairness of seven models. Four of these are baseline classifiers, chosen based on their modelling approaches (linear vs. non-linear) and their model capacities. The other 3 models were Graph Neural Networks with different aggregation mechanisms. A fuller explanation of the choice of these models is provided below. At each fold of cross-validation, the models were trained on 80% of the data and tested on the other 20%. Both the overall accuracy of the model and their fairness, measured through their performance with respect to the African–American group, was measured at each step. The aggregate of these results from the 5-fold cross-validation was then used to determine which classifier was the most accurate as well as the fairest.

For accuracy, each model was evaluated using binary accuracy, the receiver operating characteristic (ROC) curve (AUC), Precision, Recall, and the F1-score. Precision is measured as *True Positives/(True Positives + False Positives)*, Recall is measured as *True Positives/(True Positives + False Negatives)*, while the F1-Score is the harmonic mean of Precision and Recall. AUC, meanwhile, is "especially useful for domains with skewed class distribution and unequal classification error costs" as it is insensitive to class distributions [16]. The larger the AUC, the larger the probability that the classifier will rank randomly chosen positive instances higher than randomly chosen negative instances. Taken together, these metrics help evaluate the trade-offs made by models between decreasing false positives and increasing false negatives. Fairness, meanwhile, is evaluated by considering the performance of each classifier on the subset of African–American users. Particular attention is given to false positive rate (FPR) across demographic groups, for the reasons described above. The binary accuracy of each classifier is also reported for the African–American user subset in order to better understand any accuracy–fairness trade-off.

The hyperparameters of each model were tuned using a grid search approach, whereby a search space is defined as a grid of hyperparameter values and the model is evaluated using hyperparameter combinations on each position in the grid [23]. The hyperparameters include the L2 regularization strength parameter for logistic regression, the regularization parameter and scaling parameter for SVM, and the learning rate, batch size, and layer number for the ANNs. For the GraphSAGE models, the hyperparameters tuned include the learning rate, aggregation dimension, and neighbourhood depth.

Finally, a detailed error analysis of the best performing model is performed to further understand areas where even the fairest and most accurate classifier may still face limitations.

4.2 Detecting African–American users using a mixed-membership demographic language model

Given the size of social media datasets, large-scale statistical tools have been developed to automatically infer demographic attributes, which can be used to diagnose and mitigate potential bias in the models deployed on such data [63]. A subset of such research is devoted to inferring the race or ethnicity of social media users based on their content, profiles, or both (e.g., [3, 50]). For this paper, we use a distantly supervised probabilistic topic model developed by Blodgett et al. [3]. It models dialect variation across races in the US [18], using Latent Dirichlet allocation (LDA) to classify users into the four main categories of the US Census: non-Hispanic white, non-Hispanic black, Hispanic, and Asian. Using this model, Blodgett et al. [3] report accuracy of 97% (AUC) for detecting users' race in a dataset of 26,009 users. It is appropriate for this research as the model was trained, at least partly, on US Census data and is likely to be most effective for US-based populations. Given that the random-walk induced strategy for creating the retweet network used here had as its starting seed node a user in San Jose, California, our sample of users is likely to mostly be from the USA [20].

Following a similar approach to Preot and Ungar [50], we applied the pre-trained topic model of [3] to the messages of each user, averaging the probabilities to determine a score for each user. If the average probability for the African–American category was above 0.8 we labelled the user as African–American. At this first step, 168 users were identified with this approach; further manual verification was conducted at a second step, after which 32 users were identified as false positives, leaving 136 African–American users in the dataset. Of these, eight are labelled as hateful.

4.3 Classifying hateful users using machine learning

In total we train seven models, four of which incorporate network information. 5-fold cross-validation is used to test the predictive power and fairness of the models, with a 80/20 train–test split. All models are coded in Python using the TensorFlow2 library.

4.3.1 Baseline classifiers

We train four baseline classifiers selected based on approaches common in computational social science research on hateful content detection [60]: regularized logistic regression, Support Vector Machine (SVM) with a radial basis function kernel, and two fully connected Artificial Neural Networks (ANNs). We compare the performance of graph neural networks (GNNs) against these baseline models on both accuracy and fairness. The logistic regression was chosen due to its linear modelling of the problem, which is contrasted with the non-linear approach of the SVM and ANNs. The choice of models also reflects a progressive increase in model capacity, with the logistic regression having the lowest capacity, followed by the SVM and the ANNs. Additionally, one of the ANNs is trained on textual, user-level, and manually engineered network-level features described above; all other baseline classifiers are trained only on textual and user-level features. Given the complexity of the problem and the superior model capacity of the ANN, it was considered likely that it would have superior performance. As such, we experimented with two different versions of the ANN: one trained on just textual and user-level data, and another with manually engineered network-level features. The ANN with network-level features was chosen both to see if such features had an incremental impact on its performance but also

to better compare this model to the GNNs, which also incorporate network information but learn network-level features in an end-to-end manner.

4.3.2 Geometric deep learning and GraphSAGE

Analysis of properties of edges, nodes and topologies of such networks has been subject to various statistical and social methodologies (see [31]). However, predicting such properties using machine learning had been restricted to using shallow representations or encodings of such properties, such as through node2vec [19] or TransE [4]. Such shallow encoders, however, cannot generalize to unseen nodes. The entire graph must be first embedded before prediction can be performed [34]. This makes them impractical for use on most online social networks, where new users are added to the network in a continual stream. Moreover, they incorporate only the graph's topology, ignoring attributes of the nodes themselves, such as a user's text or activity patterns in the current case [34].

Geometric deep learning, through the use of GNNs, overcomes such difficulties, resulting in more information-rich and predictive node embeddings fine-tuned to a given task [67]. Spatial Graph Convolutional Neural Networks (GCNNs), a variant of which is used here, create node embeddings by defining graph convolutions upon a node's spatial relations: each node's representations are obtained by convolving its embedding with those of its neighbours, leading to information propagation across a graph's edges while capturing its topological structure simultaneously [67]. A spatial approach was chosen, instead of the also popular spectral approach, due to spatial approaches' better efficiency, more effective generalisability to new graphs, and capacity to incorporate directed edges [67]. The latter two properties are especially relevant for constantly evolving and directed social networks such as Twitter.

We use GraphSAGE, an inductive spatial variant of GCNNs proposed by Hamilton et al. [21]. As opposed to learning distinct embeddings for each node, in what would be a 'transductive' approach, GraphSAGE trains "a set of aggregator functions that learn to aggregate feature information from a node's local neighbourhood", with each such function aggregating information from a different number of hops away from the target node in the network [21]. The learnt parameters of these functions, shared across nodes in a form analogous to vanilla convolutional neural networks, are then used to generate embeddings for previously unseen nodes. Such an 'inductive' or end-to-end approach learns both the topological structure of node neighbourhoods as well as the optimal aggregation of node features within them [21].

GraphSAGE can incorporate different aggregation functions, and we use three GraphSAGE models with different aggregation functions: mean aggregation, which takes a weighted average of the embeddings of a node's neighbours; max pooling, which transforms these embeddings through a neural network and takes the maximum activation; and self-attention, which uses a single attention layer to compute attention coefficients for a node's neighbours, better incorporating their differing importance for the target node's representation.

Another popular kind of GNN, Graph Convolutional Networks (GCNs), are not used because of their high memory requirements and computational complexity. When training GCNs, usually the whole graph data and intermediate states of all nodes is required to fit into memory [67]. In contrast, GraphSAGE recursively samples a node's neighbourhood for a fixed K number of steps, providing a more memory efficient solution as the whole

Table 1 Comparison of Complexity of GraphSAGE and Graph Convolutional Networks (GCN) (Adapted from [67]) n is the total number of nodes. m is the total number of edges. K is the number of layers. s is the batch size. r is the number of neighbours being sampled for each node

Complexity	GCN	GraphSAGE
Time	$O(Kmd + Knd^2)$	$O(r^K nd^2)$
Memory	$O(Knd + Kd^2)$	$O(sr^K d + Kd^2)$

Table 2 Performance metrics ± standard deviations (Maximum mean columnar values in bold)

Model	Accuracy	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
LR	83.1 ± 1.78	37.6 ± 1.16	82.1 ± 1.32	51.6 ± 1.54	89.9 ± 1.20
SVM	81.6 ± 0.89	34.8 ± 0.77	77.6 ± 0.13	48.0 ± 0.64	87.6 ± 0.41
$ANN_{\text{text+user}}$	84.1 ± 2.58	40.6 ± 2.39	75.7 ± 2.14	51.7 ± 3.21	88.2 ± 2.45
$ANN_{\text{text+user+network}}$	87.2 ± 2.02	45.3 ± 2.47	68.8 ± 2.05	54.2 ± 3.08	87.2 ± 1.72
$GraphSAGE_{\text{maxpool}}$	84.2 ± 1.59	40.3 ± 1.46	80.9 ± 1.08	53.4 ± 1.16	90.2 ± 1.34
$GraphSAGE_{\text{attention}}$	**84.8** ± 1.11	40.6 ± 1.25	**82.3** ± 1.18	54.3 ± 1.14	90.8 ± 1.07
$GraphSAGE_{\text{meanagg}}$	87.4 ± 1.04	**46.1** ± 1.15	76.8 ± 1.11	**57.5** ± 1.12	**90.8** ± 1.03

graph is no longer needed [67]. GCNs, moreover have a computational time complexity of $O(m)$, where m represents the number of edges in the graph [67]. GraphSAGE, meanwhile, has a time complexity dependent on the number of samples and layers; these can be tuned to make GraphSAGE more efficient [67]. Lastly, as Hamilton et al. [21] demonstrate, on a social network node classification task similar to the one used here, GraphSAGE generates a higher F1 score than GCN, despite its lower computational and memory burden. We also experimented with GCN on an NVIDIA Tesla M60 GPU with 2048 parallel processing cores, but found that it was too time-consuming on the present network of +2M edges. A fuller comparison of GraphSAGE and GCN's memory and time complexity is provided in Table 1. Given these reasons, we did not employ GCNs, but instead experimented with GraphSAGE and its different aggregation methods.

5 Results

Table 2 shows the performance of each model. In terms of binary accuracy as well as F1-Score, models that incorporate network-level features perform better than those that do not, while neural-network based approaches outperform more traditional machine learning approaches. Geometric deep learning performs best across all metrics. The GraphSAGE model (using mean aggregation) has the best F1-Score (57.5) and AUC (90.8).

Our results show the power of geometric deep learning for achieving higher performance in hateful user detection. Firstly, the logistic regression model, the simplest baseline used, still performs fairly well, even surpassing the ANNs on certain metrics. Secondly, when incorporating network information into the ANN, its performance declines considerably on some metrics such as Recall (–6.9) and AUC (–1). However, these declines do not occur for the GraphSAGE models, which also incorporate network-related information. Lastly, all models have relatively high recall but low precision, indicating that they can detect hateful users well but often mistake normal users for hateful ones. The fact that AUC is still high, despite this, indicates that this disparity may not be entirely related to trade-offs made by the models between true positives and false positives. Instead, the size and unbalanced nature of the dataset is likely to have played a role (despite our use of class weights). Indeed, as Juba and Le [28] report through an investigation of 4 imbalanced datasets, "in the presence of severe class imbalance, it is not possible to reliably achieve

Table 3 Fairness Evaluation Metrics (African–American denoted as AA, False positive denoted as FP)

Model	Accuracy (AA users)	FPs (AA users)	FPs (non-AA users)	FP rate (AA users)	FP rate (non-AA users)
$ANN_{\text{text+user}}$	80.9%	26	660	20.3%	14.8%
$ANN_{\text{text+user+network}}$	86.0%	13	**465**	10.2%	**10.5%**
$GraphSAGE_{\text{maxpool}}$	91.7%	5	680	3.9%	15.7%
$GraphSAGE_{\text{attention}}$	93.9%	1	662	0.7%	15.3%
$GraphSAGE_{\text{meanagg}}$	**94.2%**	**0**	492	**0.0%**	11.1%

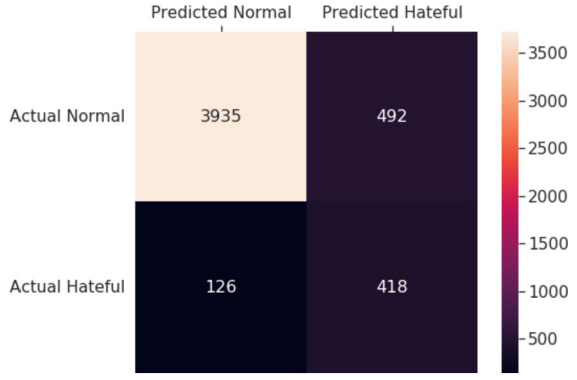

Figure 1 Confusion matrix. Predictions from best performing model: GraphSAGE with mean aggregation

high precision and recall unless one possesses a large amount of training data". Methods such as class weighting, over-sampling, or under-sampling are minimally effective and face an "inherent statistical barrier" [28].

Table 3 shows the accuracy of each model on the subset of 136 African–American users, and compares the false positive rates (FPRs) for African–American and non-African–American users. The results show that, firstly, geometric deep learning outperforms the neural network classifiers in producing the lowest FPR among African–American users, with the mean aggregation based GraphSAGE model bringing this rate to zero. Secondly, the neural network without network information ($\text{ANN}_{\text{text+user}}$) is biased against African–American users, producing a FPR that is almost 6 percentage points higher among these users compared to the non-African–American group. Incorporation of network information in the ANN helps lower and roughly equalize these rates. When such network information is dynamically learnt through the GraphSAGE models, the FPR for African–American users drops even further, becoming much lower than that of the non-African–American group.

6 Error analysis of the best performing model

Figure 1 shows the classification confusion matrix for the best performing model, GraphSAGE with mean aggregation. As shown by the performance metrics reported above, the model performs well in detecting hateful users but has problems in distinguishing normal users from hateful ones. This is seen in the larger number of false positives in the matrix's upper right corner, compared to false negatives in its lower left corner.

We conduct an in-depth error analysis to further explore the nature of the model's misclassifications. For both false positives and false negatives, we first identify quantitative differences in user and network-level features between correctly classified users and their

misclassified counterparts for the whole dataset. In a second step, in a similar vein to Vidgen et al. [59], we use the inductive, data-derived qualitative approach of grounded theory to create a typology of misclassified users. Grounded theory is well suited for error analysis as it induces themes from the data, organizes them into categories, and iterates over this process until a 'saturated' set of mutually exclusive and collectively exhaustive categories is achieved. For this qualitative step, we select a random sample of 20 users for both false positives and false negatives, and all their tweets (up to 200 each) are analysed. This is done for the entire dataset, as well as the entire subset of African–American users, reflecting the focus of each research question.

6.1 False positives

Normal users who are wrongly classified as hateful (false positives) comprise the majority of errors (∼80%). Out of these, 23 (∼5%) were suspended by Twitter a few months after the dataset's creation, indicating that there may have been an annotation error where actually hateful users were labelled as normal. However, we cannot verify this.

Interestingly, in many regards users in the false positive category are more similar to correctly classified hateful users, and dissimilar to correctly classified normal users. This may explain why the GraphSAGE model made errors in its classifications. Firstly, in terms of their social network: only 3% of users in the correctly classified normal user group (true negatives) had hateful neighbours but more than 50% of users in the false positive group did. The GraphSAGE models, which rely on the networks of users, may have failed to adequately distinguish between users whose immediate networks contained a high proportion of hateful users. This indicates that the model still encodes some 'guilt by association' but this is likely far less than with other methods. Secondly, the average number of words that users in the false positive category use from ADL and Hatebase's hateful lexicon is around 16% higher than correctly classified normal users. Lastly, users in the false positive category display more negative sentiment in their tweets than even hateful users. However, there are no meaningful differences observed in other network or user-level categories such as degree distributions or activity metrics. This analysis indicates that on a network and textual level, false positive users are more similar to correctly classified hateful users than to the correctly classified normal ones.

Using a grounded-theory approach, qualitative inspection of the false positives shows they fall into four categories: annotation errors, ambiguous content, offensive (but not hateful) content, and negative sentiment against minorities. In no cases did the GraphSAGE model make a mistake that is entirely unexplainable.

Annotation errors These are cases where annotators have made errors and as such are not attributable to the machine learning models. Five users (25% of the false positives we reviewed) fall into this category. They include both explicit and more subtle forms of hate. For example, one user in this category clearly expressed hate against Jews with the statement: "Because Jewish perverts like Weinstein, Woody Allen and Roman Polanski aren't unique, they're a trend". Others, however, express hate in more subtle ways, such as by using uncommon hateful slurs. For example, two misogynistic users often claimed that victims of rape deserved it and used terms such as 'soyboy', a deeply misogynistic term used in Men's Liberation Movement groups but relatively rare in the overall online hate landscape [27]. Moreover, four of these five users expressed irritation with Twitter about

moderating their tweets, for example: "Every post I make is marked sensitive. It is Twitter's soft censorship"

Ambiguous content This category includes errors where it is unclear whether hate was expressed. Some users are difficult for even a human to categorize, either because further context is needed to interpret their content or because the remarks are genuinely indeterminate. Examples of such content include reference to a 'slut culture problem', attributing something to the 'black race' where context is not present on what is being attributed, and implying connections between Muslims and terrorism. Further context is often needed in such cases to establish their hatefulness. Note that some of these 'ambiguous' cases could easily be considered annotation errors by some viewers (i.e., they could be genuinely hateful); however, on balance we believe there is insufficient evidence to make this stronger claim.

Offence vs. hate Around 18% of false positive users tweeted in an offensive but not overtly hateful manner, although the line between these two is itself often hard to establish [12]. All such users had a considerable number of tweets with offensive words or phrases such as 'cunts', 'pussies', 'whore', 'fuck off', or 'faggots'. Some tweeted about minority groups in an offensive, though not explicitly hateful way. For example: "Being ethnic does not stop you going to Sandhurst, being as thick as pig poo does. Positive discrimination is dumbing down" or "My ex boss would not employ women or coloureds because he thought they were lazy".

Negative sentiment but not hate against minorities Almost 30% of all false positives expressed negative sentiment but not hate against minorities, and were possibly misclassified by the model partly due to this factor. LGBT groups, immigrants, and Muslims feature prominently in such cases. Examples include: "She's illegal you dipshit. A felon. She has no rights in this country, nor do you in hers. Gtfoh" or "Liberals: Lets use Islamic teachings to protect women. Islam: Lets make women property. Weird world we live in".

6.2 False negatives

Hateful users who are wrongly classified as normal (false negatives) comprise the minority of errors (∼20%). In contrast with the false positives, these users are not highly dissimilar to users from their true class (here, hateful). That said, they exhibit some notable similarities with users in their wrongly predicted class (i.e., normal), which may partially explain why the model misclassified them. A far lower proportion of false negatives had normal neighbours compared with an average hateful user: on average, around 80% of correctly classified hateful users had a normal neighbour, compared to only 51% of false negative users.

Even on a textual level, the average number of hateful words used by the hateful users (true positives) and the false negatives is nearly equal, and 37.5% higher than the average for normal users. Yet, in other areas such as sentiment and subjectivity, false negatives are systematically dissimilar to hateful users. False negatives show, on average, 25% greater positive sentiment in their tweets versus hateful users.

Qualitative analysis of false negative users reveals three major categories: annotation errors, hate against uncommon targets, and non-explicit hate.

Annotation errors A large portion of false negatives (40%) were due to annotation errors. Firstly, hateful labels were often wrongly assigned to users who were, in fact, opposed to hateful groups and used strongly-worded language to defend minority groups. For example, a user who was against the genocide of Rohingya Muslims in Myanmar and expressed vigorous opposition to its regime was annotated as hateful. Secondly, a quarter of these wrongly annotated users used either obscene or offensive language, but in self-referential and non-hateful ways. Lastly, a third of these annotation errors were for African–American users. They showed frequent use of words such as 'nigga', 'bitch', or 'ass' but always in non-hateful ways and likely in community contexts. Such errors reflect previous findings that even annotators focus strongly on words only, often ignoring the context of language use and thus mislabelling users as hateful [45]. Note that no such errors were identified with African–American users for the false positives.

Hate against uncommon targets Of the remaining 60% of false negatives that can be directly attributed to the model, a quarter contain users who express hate against uncommon targets. Such users were indeed hateful, but directed their hate against Native-Americans, old-aged people, or Christians. These categories of victims are uncommon on social media: research by Mondal et al. [44] on categorizing hate speech on Twitter by its targets finds that hate directed against religion or ethnicity make up less than 2% of all hate speech. Hate against old people does not feature at all in their data. Given the scarcity of such examples, ML models may not adequately learn to identify such forms of niche hate.

Implicit hate A third of the model-attributable false negatives were for implicit hate: hateful users who did not explicitly use hateful words or phrases in their texts but relied on contextual references. For example, two users praised ISIS's actions and someone who "beat up the black guy on the bus in that video" respectively. Another favourably referenced what Hitler would do in a particular situation to express racist sentiment. This category also featured users who use subtle but still hateful words when diminishing women, such as 'broad'.

6.3 Errors among African–American users

For African–American users the GraphSAGE model only produced false negatives, with no false positives. 50% of these were annotation errors where words, phrases, and other linguistic features common to AAVE were likely mistaken by annotators as signals for the hateful nature of a user, a source of annotation error common in many other popular hate detection datasets (see [14, 55]). Users with tweets such as "That's flint nigga not Detroit. Dumbass" or "God damn your bitch ass backup account do numbers like this?" were mislabelled as hateful. From an analysis of all their available tweets, however, it is clear that they use offensive or otherwise hateful terms such as 'nigga' or 'bitch' in community contexts only, without any overt intention of harming a potential target. Out of the other half, which comprised errors produced by the model, one user had been suspended from Twitter. They expressed anti-LGBT hate, albeit more subtle varieties. The others either had content that could be characterized as offensive, for example "I like nasty bitches", or expressed hate implicitly.

7 Discussion

In this section, we discuss the results in the context of our objective of producing more accurate and fairer hateful user classifications. We divide the discussion into two subparts: Sect. 7.1 focuses on the accuracy of our proposed system while Sect. 7.2 examines fairness. The strengths and limitations of the system are discussed, and some potential future research areas are highlighted.

7.1 Towards more accurate classification

Our results show that the incorporation of dynamically learnt network representations along with user-level and textual information through geometric deep learning produces more accurate classification of hateful users. Across all metrics, GraphSAGE performs the best. From a social science perspective, our results help support existing theories about homophily among hateful users by demonstrating how network-level insights can be leveraged successfully to boost classification accuracy. The most accurate results are only achieved when users' network-related information is dynamically learnt. The recall for the ANN incorporating static network features is actually the lowest among all models. We showed that normal users often had a considerable proportion of hateful neighbours. Even so, models including network information have the highest precision scores, indicating that they are best able to distinguish between the two classes of users by using their networks. Leveraging such features more efficiently and effectively could demonstrably improve flagging of hateful users in the real world.

Despite the advances that geometric deep learning offers, precision remains fairly low overall, with the best classifier having a score of 46.1. While this may be because of the nature of the retweet network, where normal users do not show particularly strong assortative mixing, it may also arise due to training data scarcity, which is a concern here given that the dataset is an order of magnitude smaller than most labelled hate speech training datasets [59]. Furthermore, as the error analysis demonstrates, annotation errors comprised a substantial proportion of both false negatives and false positives and could also contribute to this, artificially depressing the model's performance. Further work should probe the underlying tensions faced by human annotators in labelling such content. As Vidgen et al. [59] write, for "complex and often-ambiguous content even well-trained annotators can make decisions [that] are inappropriate". Disagreement metrics between annotators, such as Kappa scores, are essential for understanding whether the model's errors are related to annotator uncertainty (i.e., the 'ground-truth' label is contested in the first place). Unfortunately, no such metrics were collected by the datasets' creators, preventing such analysis in this case.

The model performs poorly when users' content is ambiguous or where hate is not explicitly expressed. It is unable to adequately distinguish between offence and hate, possibly because it focuses on hateful words without fully understanding the nuanced context in which they are used. This is partly because of the nature of the data itself. Tweets are short-form content, often implicitly reference other entities, and can be ambiguous, even when considered by humans [37]. A way to ameliorate this problem may be through the consideration of heterogeneous graphs, where users and their tweets are considered as heterogeneous but connected nodes, and different types of edges are placed between users and between their tweets if a tweet references another one (for example, through a retweet or reply). Indeed early experiments by Facebook AI in using heterogeneous graphs

combined with graph convolutional networks for hateful *content* detection have shown promising results [42]. Another way to help address this problem may be by using a better representation of each user's content. We used GloVe embeddings (provided by the original authors) but contextual word embeddings, such as BERT, have been shown to achieve better performance in classifying hateful content [45].

7.2 Geometric deep learning: A fairer approach?

Alongside boosting accuracy, we showed that incorporating learnt network representations of users into the classification task also reduces longstanding biases against African–Americans in automated hate detection. This boosts fairness as far as 'predictive equality' is concerned, where non-discrimination towards the minority group is required only for a particular outcome of interest, which in this context is being classified as hateful. Without the inclusion of network information, the ANN classifier produced disparate impact on African–American users, who had an FPR almost 6 percentage points higher than users from other demographic groups. The inclusion of network information within the ANN itself helped somewhat to equalize the FPRs between the minority and majority groups. When such network representations are learnt end-to-end, the FPR declines further for the African–American group, dropping to zero for GraphSAGE with mean aggregation, which is also the most accurate classifier overall. Thus, it appears that the selection of the 'predictive equality' fairness criteria, which unlike other criteria such as 'demographic parity' does not place strict limits on classification accuracy, has the desired effect: the fairest classifier is also the most accurate one.

These results, however, should be interpreted with some caution. Firstly, the subset of African–American users identified is small, and testing on a larger group is required to fully establish the fairer nature of geometric deep learning classifiers. Secondly, the annotation errors observed in the error analysis complicate what conclusions can be drawn from the results. Half of the 'false-negatives' among African–American were users labelled as hateful despite being, in fact, normal.

Additionally, geometric deep learning methods decrease the FPR among African–American users to a rate that is *lower* than that of the non-African–American group. From a purely technical perspective, where 'predictive equality' is concerned, the FPR between the two groups should be close to, if not exactly, equal. The large gap in the FPR rates, despite addressing historical bias against a minority group, thus poses a problem from this viewpoint. From a political and moral philosophy perspective, judgements on fairness are often driven by broader normative principles [33]. In this regard, an unequal error rate could be used to correct a past bias. For instance, hate detection algorithms have been shown to be biased against African–Americans, persistently suppressing their speech unfairly. To remedy this, a normative decision could be made to assign greater importance to lowering the FPR for this group than for the non-African–American group. This could be considered more reflective of societal ideals of fairness.

8 Conclusion

We have demonstrated the promise of using geometric deep learning to incorporate network-level information into hateful user classification, showing that it can lead to fairer and more accurate results. This is one of the first works (to our knowledge) to both consider the problem of fairness in hateful user detection and to develop a model that mini-

mizes it while also boosting accuracy. By leveraging the rich context that network information provides, our system dealt better with the problems of polysemy that have often led automated hate detection to be inaccurate and biased. It allowed our system to distinguish between hateful and non-hateful users who otherwise shared many linguistic characteristics.

In completing this study we have leveraged both social science theories and recent advances in ML research. The task of automated hate detection is, despite the widespread use of ML models, an inherently social one. Combining social science insights on homophily between hateful users with geometric learning techniques, this paper has enabled creation of a more effective classifier. It has evaluated this classifier, moreover, by choosing an appropriate notion of fairness that is guided by broader concerns in society about historic algorithmic bias against African–Americans (predictive equality).

With this contribution, we hope to inspire future work in this nascent but important area of research. A fruitful area for future research would be to explore further ways and situations in which network-information can enhance automated hate detection, testing the generalisability of such methods to more contexts.

Acknowledgements
Not applicable.

Funding
This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Criminal Justice System" theme within that grant, and by The Alan Turing Institute.

Abbreviations
AA, African–American; AAVE, African–American Vernacular English; ANN, Artificial Neural Network; AUC, Area under the ROC Curve; DURW, Directed Unbiased Random Walk; FPR, False Positive Rate; GCN, Graph Convolutional Network; GNN, Graph Neural Network; LDA, Latent Dirichlet allocation; LR, Logistic Regression; ML, Machine Learning; SAE, Standard American English; SVM, Support Vector Machine; ROC, curve: receiver operating characteristic curve.

Availability of data and materials
The datasets analysed during the current study are created by Ribeiro et al. [53] and available at `https://github.com/manoelhortaribeiro/HatefulUsersTwitter`

Declarations

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors designed the study. ZA collected the data, performed the analysis, and wrote the first draft. BV and SAH revised the draft. All authors gave final approval for publication.

Author details
[1]Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK. [2]Alan Turing Institute, 2QR, 96 Euston Road, London, NW1 2DB, UK. [3]Meedan, 160 Spear St., Suite 1000, San Francisco, CA 94105, USA.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 10 May 2021 Accepted: 24 January 2022 Published online: 14 February 2022

References

1. Alorainy W, Burnap P, Liu H, Williams ML (2019) "The enemy among us": detecting cyber hate speech with threats-based othering language embeddings. ACM Trans Web 13(3). https://doi.org/10.1145/3324997
2. Barocas S, Hardt M (2017) NIPS 2017 Tutorial: fairness in machine learning. https://arxiv.org/abs/2005.03909
3. Blodgett SL, Green L, O'Connor B (2016) Demographic dialectal variation in social media: a case study of African–American English. In: EMNLP 2016—conference on empirical methods in natural language processing, proceedings. https://doi.org/10.18653/v1/d16-1120

4. Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation
5. Bowman-Grieve L (2009) Exploring stormfront: a virtual community of the radical right. Stud Confl Terrorism 32(11):989–1007. https://doi.org/10.1080/10576100903259951
6. Burnap P, Williams ML (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. https://doi.org/10.1002/poi3.85
7. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Mean birds: detecting aggression and bullying on Twitter. In: WebSci 2017—proceedings of the 2017 ACM web science conference,association for computing machinery, pp 13–22. https://doi.org/10.1145/3091478.3091487
8. Chiril P, Pamungkas EW, Benamara F, Moriceau V, Patti V (2022) Emotionally informed hate speech detection: a multi-target perspective. Cogn Comput 14:322–352
9. Christ O, Schmid K, Lolliot S, Swart H, Stolle D, Tausch N, Al Ramiah A, Wagner U, Vertovec S, Hewstone M (2014) Contextual effect of positive intergroup contact on outgroup prejudice. In: Proceedings of the National Academy of Sciences of the United States of America 111. https://doi.org/10.1073/pnas.1320901111
10. Chung A (2019) How automated tools discriminate against black language. https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db
11. Davidson T, Bhattacharya D, Weber I (2019) Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the third workshop on abusive language online, association for computational linguistics, Florence, Italy, pp 25–35. https://doi.org/10.18653/v1/W19-3504
12. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th international conference on web and social media, ICWSM 2017
13. Day E (2018) The application of machine learning, big data techniques, and criminology to the analysis of racist tweets. PhD thesis, Canterbury Christ Church University
14. Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. In: AIES 2018—proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society. https://doi.org/10.1145/3278721.3278729
15. Eckstrand N (2018) The ugliness of trolls: comparing the methodologies of the alt-right and the Ku Klux Klan. Cosmopolitan civil societies. Interdiscip J 10:41–62. https://doi.org/10.5130/ccs.v10.i3.6026
16. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8). https://doi.org/10.1016/j.patrec.2005.10.010
17. Fehn Unsvåg E, Gambäck B (2018) The effects of user features on Twitter hate speech detection. In: Proceedings of the 2nd workshop on abusive language online (ALW2), association for computational linguistics, Brussels, Belgium, pp 75–85. https://doi.org/10.18653/v1/W18-5110
18. Green LJ (2002) African American English. Cambridge University Press, Cambridge. https://doi.org/10.1017/cbo9780511800306
19. Grover A, Leskovec J (2016) Node2vec: scalable feature learning for networks. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol 13, 17-August-2016. https://doi.org/10.1145/2939672.2939754
20. Hale SA (2014) Global connectivity and multilinguals in the Twitter network. In: Proceedings of the SIGCHI conference on human factors in computing systems CHI'14. ACM, New York, pp 833–842. https://doi.org/10.1145/2556288.2557203
21. Hamilton WL, Ying R, Leskovec J (2017) Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, vol 2017-December
22. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. arXiv e-prints 1610.02413
23. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer series in statistics. Springer, New York
24. Irani D, Wrat A, Amir S (2021) Early detection of online hate speech spreaders with learned user representations. In: Faggioli G, Ferro N, Joly A, Maistro M, Piroi F (eds) Proceedings of the working notes of CLEF 2021—conference and labs of the evaluation forum, Bucharest, Romania, September 21st-to-24th. CEUR workshop proceedings, vol 2936. CEUR-WS.org, pp 2004–2010. http://ceur-ws.org/Vol-2936/paper-174.pdf
25. Jia J, Wang B, Zhang L, Gong NZ (2017) Attriinfer: inferring user attributes in online social networks using Markov random fields. In: Proceedings of the 26th international conference on world wide web, international world wide web conferences steering committee, Republic and Canton of Geneva, CHE, WWW'17, pp 1561–1569. https://doi.org/10.1145/3038912.3052695
26. Johnson NF, Leahy R, Restrepo NJ, Velasquez N, Zheng M, Manrique P, Devkota P, Wuchty S (2019) Hidden resilience and adaptive dynamics of the global online hate ecology. https://doi.org/10.1038/s41586-019-1494-7
27. Jones C, Trott V, Wright S (2019) Sluts and soyboys: mgtow and the production of misogynistic online harassment. New Media Soc 19. https://doi.org/10.1177/1461444819887141
28. Juba B, Le HS (2019) Precision-recall versus accuracy and the role of large data sets. In: Proceedings of the AAAI conference on artificial intelligence, association for the advancement of artificial intelligence, pp 4039–4048
29. Kiela D, Firooz H, Mohan A, Goswami V, Singh A, Ringshia P, Testuggine D (2020) The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. 2005.04790
30. Kleinberg JM, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. CoRR 1609.05807
31. Kolaczyk ED (2009) Statistical analysis of network data: methods and models, 1st edn. Springer, Berlin
32. Lahoti P, Beutel A, Chen J, Lee K, Prost F, Thain N, Wang X, Chi EH (2020) Fairness without demographics through adversarially reweighted learning. arXiv e-prints arXiv:2006.13114
33. Leben D (2020) Normative principles for evaluating fairness in machine learning. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society, AIES'20 Association for Computing Machinery, New York, pp 86–92. https://doi.org/10.1145/3375627.3375808
34. Leskovec J (2019). Lecture: graph neural networks

35. Lewis K, Gonzalez M, Kaufman J (2012) Social selection and peer influence in an online social network. Proc Natl Acad Sci 109(1):68–72. https://doi.org/10.1073/pnas.1109739109
36. Liu P, Li W, Zou L (2019) NULI at SemEval-2019 task 6: transfer learning for offensive language detection using bidirectional transformers. In: Proceedings of the 13th international workshop on semantic evaluation, association for computational linguistics, Minneapolis, Minnesota, USA, pp 87–91. https://doi.org/10.18653/v1/S19-2011
37. MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: challenges and solutions. PLoS ONE 14(8):1–16. https://doi.org/10.1371/journal.pone.0221152
38. Marwick A, Reidenberg JR, Russell NC (2014) Online harassment, defamation, and hateful speech: a primer of the legal landscape clip study advisors. Fordham Center on Law and Information Policy Report (July 2013)
39. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A survey on bias and fairness in machine learning. arXiv e-prints arXiv:1908.09635
40. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations ofwords and phrases and their compositionality. In: Advances in Neural Information Processing Systems
41. Mishra P, Del Tredici M, Yannakoudakis H, Shutova E (2018) Author profiling for abuse detection. In: Proceedings of the 27th international conference on computational linguistics, association for computational linguistics, Santa Fe, New Mexico, USA, pp 1088–1098. https://www.aclweb.org/anthology/C18-1093
42. Mishra P, Tredici MD, Yannakoudakis H, Shutova E (2019) Abusive language detection with graph convolutional networks. CoRR 1904.04073
43. Mishra S, Prasad S, Mishra S (2021) Exploring multi-task multi-lingual learning of transformer models for hate speech and offensive speech identification in social media. SN Comput Sci 2:72. https://doi.org/10.1007/s42979-021-00455-5
44. Mondal M, Silva LA, Benevenuto F (2017) A measurement study of hate speech in social media. In: Proceedings of the 28th ACM conference on hypertext and social media, association for computing machinery, HT'17. Association for Computing Machinery, New York, pp 85–94. https://doi.org/10.1145/3078714.3078723
45. Mozafari M, Farahbakhsh R, Crespi N (2019) A BERT-based transfer learning approach for hate speech detection in online social media. arXiv e-prints arXiv:1910.12574
46. Mullah NS, Zainon WMNW (2021) Advances in machine learning algorithms for hate speech detection in social media: a review. IEEE Access 9(88):364–388. https://doi.org/10.1109/ACCESS.2021.3089515
47. Olteanu A, Talamadupula K, Varshney KR (2017) The limits of abstract evaluation metrics: the case of hate speech detection. In: Proceedings of the 2017 ACM on web science conference. Association for Computing Machinery, New York, pp 405–406. https://doi.org/10.1145/3091478.3098871
48. Pavlopoulos J, Sorensen J, Dixon L, Thain N, Androutsopoulos I (2020) Toxicity detection: does context really matter? arXiv e-prints arXiv:2006.00998
49. Pitsilis GK, Ramampiaro H, Langseth H (2018) Detecting offensive language in tweets using deep learning. arXiv e-prints 1801.04433
50. Preot D, Ungar L (2018) User-level race and ethnicity predictors from Twitter text. In: Proceedings of the 27th international conference on computational linguistics
51. Rangel F, Sarracén G, Chulvi B, Fersini E, Rosso P (2021) Profiling hate speech spreaders on Twitter task at pan 2021. In: Faggioli G, Ferro N, Joly A, Maistro M, Piroi F (eds) Proceedings of the working notes of CLEF 2021—conference and labs of the evaluation forum. Bucharest, Romania, September 21st-to-24th. CEUR workshop proceedings, vol 2936. CEUR-WS.org, pp 2004–2010. http://ceur-ws.org/Vol-2936/paper-149.pdf
52. Ribeiro B, Wang P, Murai F, Towsley D (2012) Sampling directed graphs with random walks. In: Proceedings—IEEE INFOCOM. https://doi.org/10.1109/INFCOM.2012.6195540
53. Ribeiro M, Calais P, Santos Y, Almeida V, Wagner M (2018) Characterizing and detecting hateful users on Twitter. arXiv e-prints 1803.08977
54. Sabatini F, Sarracino F (2017) Online networks and subjective well-being. Kyklos 70(3):456–480. https://doi.org/10.1111/kykl.12145
55. Sap M, Card D, Gabriel S, Choi Y, Smith NA (2019) The risk of racial bias in hate speech detection. https://doi.org/10.18653/v1/p19-1163
56. Statt N (2017) YouTube is facing a full-scale advertising boycott over hate speech
57. Tran T, Hu Y, Hu C, Yen K, Tan F, Lee K, Park SR (2020) HABERTOR: an efficient and effective deep hatespeech detector. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), association for computational linguistics, online, pp 7486–7502. https://doi.org/10.18653/v1/2020.emnlp-main.606
58. Verma S, Rubin J (2018) Fairness definitions explained. In: Proceedings of the international workshop on software fairness, FairWare'18. Association for Computing Machinery, New York, pp 1–7. https://doi.org/10.1145/3194770.3194776
59. Vidgen B, Hale S, Guest E, Margetts H, Broniatowski D, Waseem Z, Botelho A, Hall M, Tromble R (2020) Detecting East Asian prejudice on social media. In: Proceedings of the fourth workshop on online abuse and harms, association for computational linguistics, online, pp 162–172. https://doi.org/10.18653/v1/2020.alw-1.19
60. Vidgen B, Harris A, Nguyen D, Tromble R, Hale S, Margetts H (2019) Challenges and frontiers in abusive content detection. In: Proceedings of the third workshop on abusive language online, association for computational linguistics, Florence, Italy, pp 80–93. https://doi.org/10.18653/v1/W19-3509
61. Vidgen B, Thrush T, Waseem Z, Kiela D (2020) Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. 2012.15761
62. Vidgen B, Yasseri T (2018) Detecting weak and strong Islamophobic hate speech on social media. arXiv e-prints arXiv:1812.10400
63. Wang Z, Hale S, Adelani DI, Grabowicz P, Hartman T, Flöck F, Jurgens D (2019) Demographic inference and representative population estimates from multilingual social media data. In: The world wide web conference, WWW'19. Association for Computing Machinery, New York, pp 2056–2067. https://doi.org/10.1145/3308558.3313684
64. Waseem Z, Hovy D (2016) Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. https://doi.org/10.18653/v1/n16-2013
65. Wick M, Panda S, Tristan JB (2019) Unlocking fairness: a trade-off revisited. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R (eds) Advances in neural information processing systems, vol 32. Curran Associates, Red Hook, pp 8783–8792. http://papers.nips.cc/paper/9082-unlocking-fairness-a-trade-off-revisited.pdf

66. Wimmer A, Lewis K (2010) Beyond and below racial homophily: Erg models of a friendship network documented on Facebook. Am J Sociol 116(2):583–642. http://www.jstor.org/stable/10.1086/653658
67. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS (2020) A comprehensive survey on graph neural networks. In: IEEE transactions on neural networks and learning systems, pp 1–21. https://doi.org/10.1109/tnnls.2020.2978386
68. Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Mubarak H, Derczynski L, Pitenis Z, Çöltekin Ç (2020) SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). arXiv preprint 2006.07235

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com