International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.165

Hate speech Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model

Dr.M.Prakash
Department of Data Science and Business System,
School of Computing
SRM Institute of Science and Technology, Kattankulathur
TamilNadu
prakashm2@srmist.edu.in

VegiYaswanth
Department of Data Science and Business System,
School of Computing
SRM Institute of Science and Technology, Kattankulathur
TamilNadu
yv5117@srmist.edu.in

NittalaSatyaSaiKireeti
Department of Data Science and Business System,
School of Computing
SRM Institute of Science and Technology, Kattankulathur
TamilNadu
kn6196@srmist.edu.in

***Abstract*—This study focuses on detecting racist content in Tweets using sentiment analysis to address the problem of hate speech on social media. A Gated Convolutional Recurrent-Neural Networks (GCR-NN) model was developed by combining gated recurrent unit (GRU), convolutional neural networks (CNN), and recurrent neural networks (RNN) to accurately detect racist comments in tweets. The GCR-NN model achieved an accuracy of 0.98, outperforming other models in identifying subtle and concealed instances of racism. The study suggests that the GCR-NN model has the potential to be a valuable tool for social media platforms in the fight against hate speech and promotion of a more inclusive online environment.**

Keywords—Gated Convolutional Recurrent-Neural Networks (GCR-NN) model, Recurrent Neural Networks(RNN), Hate Speech Detection, Convolutional neural networks(CNN)

I. INTRODUCTION

Social media has evolved into a powerful force in the socio-political landscape, influencing our behavior and shaping our perspectives in numerous ways. However, with the widespread use of social media platforms and the freedom of expression that they offer, a multitude of negative aspects have emerged in recent years, including the proliferation of hate speech and racism. Social media has revolutionized the way we interact with the world around us, shaping our thoughts and actions in numerous ways. However, with the widespread use of social media platforms and the freedom of speech they afford, several negative phenomena have emerged in recent years, including hate speech and racism.

Twitter, in particular, has become a new setting in which racism and related stress seem to thrive. Currently, 22% of adults in the United States use Twitter, and the platform boasts a staggering 1.3 billion accounts and 336 million active users worldwide, generating around 500 million tweets per day. Unless users opt to make their tweets private, they are publicly available and can be reacted to, shared, tagged, liked, or responded to by other Twitter users.

The growing popularity of social media platforms has led to their wide use for several old and new forms of racist practices. Hate is expressed on such platforms in different surreptitious forms such as memes and openly such as posting Tweets containing racist remarks using fake identities. Although often associated with ethnicity, hate is now thriving based on color, origin, language, cultures, and most importantly religion. Social media opinions and remarks provocating racial differences have been regarded as a serious threat to social, political, and cultural stability and have threatened the peace of different countries. Social media being the leading source of hate opinions dissemination should be monitored and hate speech remarks should be detected and blocked timely.

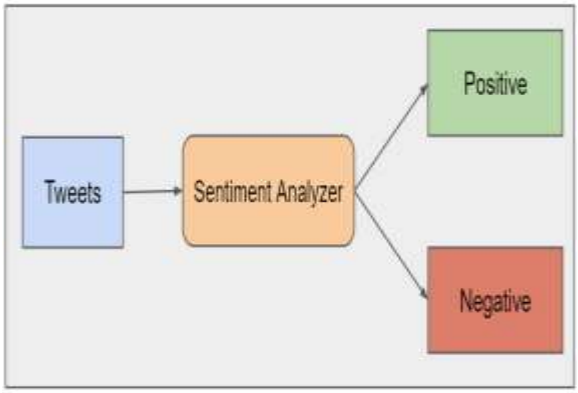

Fig.1: Example figure

Hate speech remarks and tweets on social media have been linked to a variety of mental and physical illnesses, resulting in negative health consequences [7–12]. Hate speech on social media may be classified into three types: institutionalized, individually mediated, and internalized [13]. Hate speech may be experienced personally via hate speech discrimination or unequal racial treatment, as well as knowledge of prejudice against relatives and friends. In today's society, the issue of hate conduct has become increasingly prevalent and has had a detrimental impact on individuals, causing a wide range of psychosocial stress and increasing the risk of chronic illnesses [14-16]. This is a concerning issue that has become amplified by the proliferation of social media platforms, where hate spreading organisations and individuals have been able to promote their agenda with increased skill and complexity through numerous channels and techniques. In response to this problem, various techniques have been developed and special attention has been given to the area of sentiment analysis in order to evaluate text from social media platforms for a wide range of tasks such as hate speech identification, sentiment-based market prediction, and racism detection, among others. With the help of these techniques, researchers can identify and monitor hateful content in social media platforms and take necessary actions to prevent its negative impact on society.

II. Literature Review

A. *Exploring the Role of Social Media in Addressing and Combating Hate Speech: A Comprehensive Analysis*

Authors: K. R. Kaiser, D. M. Kaiser, R. M. Kaiser, and A. M. Rackham

Abstract: Social media, including sites such as Facebook, Twitter and Instagram, provides a platform for racist ideology, making this dysfunction of American society more evident. Social media can provide insight into the world of the racist-individuals who cling to their tribal identities, irrationally rejecting those who they perceive as different. Studying social media may provide insight into processes that can assist in healing American society of its segregationist views-a way toward healing the racist. The purpose of this paper is to analyse social media posts to better understand hate speech ,its causality and to develop initial steps for addressing racist ideology. To understand the behavior and mindset of American Facebook users, a comprehensive qualitative review was conducted. The study aimed to analyze the cognitive patterns, problem-solving skills, personality structures, belief systems, and coping styles exhibited by the users in their posts. The sample size for this study consisted of 600 American Facebook posts, chosen from a variety of users, age groups, and demographics. The content analysis consists of both a descriptive account of the data and an interpretive analysis. Rackham, A. M. (2018). Using social media to understand and guide the treatment of racist ideology.

B. *Methodological and Ethical Considerations for Conducting Health Research Through Social Media: Insights and Recommendations:*

As social media platforms grow in popularity and variety, so does their value for health research. Using social media to recruit participants for clinical research and/or offer health behaviour interventions may allow you to reach a larger audience. However, evidence supporting the effectiveness of these techniques is scarce, and fundamental concerns like optimum benchmarks, intervention development and methodology, participant participation, informed permission, privacy, and data management remain unanswered. Researchers interested in utilising social media for health research have little methodological advice. We outline the content of the 2017 Society for Behavioral Medicine Pre-Conference Course titled 'Using Social Media for Study,' during which the authors shared their experiences with methodological and ethical challenges related to social media-enabled research recruitment and intervention delivery. We highlight frequent problems and provide advice for social media recruiting and intervention. We also explore the ethical and appropriate use of social media in research for each of these reasons.

C. *"Cyberracism Unveiled: A Comprehensive Review of 10 Years of Research on Online Networks of Racial Hate"*

Authors: A.-M. Bliuc, N. Faulkner, A. Jakubowicz, and C. McGarty,

Abstract: The ways in which the Internet can facilitate the expression and spread of racist views and ideologies have been the subject of a growing body of research across disciplines. To date, however, there has been no systematic reviews of this research. To synthesize current knowledge on the topic and identify directions for future research, we systematically review a decade of research on cyber-racism as perpetrated by groups and individuals (i.e., according to the source of cyber-racism). Overall, the cyber-racism research reviewed shows that racist groups and individuals use different communication channels, are driven by different goals, adopt different strategies, and the effects of their communication are distinctive. Despite these differences, both groups and individuals share a high level of skill and sophistication when expressing cyber-racism. Most of the studies reviewed relied on qualitative analyses of online textual data. Our review suggests there is a need for researchers to employ a broader array of methods, devote more attention to targets' perspectives, and extend their focus by exploring issues such as the roles of Internet in mobilizing isolated racist individuals and in enabling ideological clustering of supporters of racist ideologies.

Online networks of racial hate, commonly known as "cyberracism," has been an increasing concern in recent years. In response, numerous research studies have been conducted to understand the extent, nature, and impact of these networks. This comprehensive review examines 10 years of research on online networks of racial hate, synthesizing findings from various studies and highlighting trends and patterns. The review covers topics such as the types of racist language used, the role of social media platforms in facilitating cyberracism, the psychological effects on victims, and the legal and ethical implications of cyberracism. The review also provides recommendations for addressing cyberracism, including educational and policy interventions, online monitoring, and legal actions. Overall, the review underscores the need for continued research and action to combat online networks of racial hate and their negative impact on individuals and society as a whole.

D. *Addressing Racial Health Disparities: Leveraging Existing Knowledge to Drive Action.*

Authors: D. Williams and L. Cooper

This document aims to provide a comprehensive review of the scientific data and highlight essential actions that need to be taken in order to eliminate racial health disparities. The review emphasizes the creation of communities of opportunity as a crucial first step in mitigating the negative effects of systematic racism. These communities should focus on providing resources for early childhood development, adopting policies to minimize childhood poverty, providing employment and income assistance options to adults, and promoting healthy housing and community circumstances. Secondly, the healthcare system must prioritize universal access to high-quality care, strengthen preventive healthcare approaches, address patients' social needs as part of healthcare delivery, and diversify the healthcare workforce to better reflect the patient population's demographic composition.

Additionally, further research is necessary to determine the most effective tactics for mobilizing political will and support to address health-related socioeconomic disparities. This will include initiatives to raise awareness of the prevalence of health inequities, build empathy and support for addressing inequities, strengthen individuals' and communities' capacity to actively participate in intervention efforts, and implement large-scale efforts to reduce racial prejudice, hate speech ideologies, and stereotypes in the larger culture that underpin policy preferences that initiate

and sustain inequities. Overall, the review presents a comprehensive approach to tackling racial health disparities that includes multiple levels of intervention, from creating communities of opportunity to transforming the larger cultural narrative around racial equity.

III. METHODOLOGY

In recent years, the dominance of social media in the sociopolitical sphere has led to the emergence of various forms of hate speech on the platform. Hate speech can be found in different forms on social media, including hidden forms like memes and open forms where individuals make hateful statements under false identities to provoke hate, violence, and societal instability.

The problem of racism and hate speech is not limited to ethnicity but also extends to color, origin, language, culture, and most crucially, religion. Hate thoughts and statements on social media inciting racial tensions are considered a serious threat to social, political, and cultural stability, as well as the peace of several nations. As such, social media, which is the major source of hate speech propagation, should be monitored closely, and hateful statements should be recognized and banned as soon as possible. To combat this issue, there is a need for more research and development of effective strategies and policies that can curb the spread of hate speech on social media platforms. Additionally, it is important to raise awareness and promote digital literacy to help individuals recognize and respond to hate speech effectively. It is only by taking action and working together that we can hope to create a more peaceful and just society for all.

Disadvantages:

1. Existing techniques cannot be identified and halted automatically to prevent future spread.
2. Poor performance.

In order to detect and identify instances of hate speech on Twitter, this project aims to employ sentiment analysis techniques. To achieve this, a deep learning approach is adopted, specifically a stacked ensemble model that combines gated recurrent units (GRU), convolutional neural networks (CNN), and recurrent neural networks (RNN). This ensemble model is referred to as Gated Convolutional Recurrent-Neural Networks (GCR-NN). The GCR-NN model is designed such that the GRU component is responsible for extracting meaningful features from the raw text, while the CNN component focuses on identifying key aspects that are then used by the RNN component to make accurate predictions. The use of this advanced deep learning model is expected to improve the accuracy and efficiency of detecting and identifying hate speech on Twitter.

Advantages

1. The GCR-NN model has been proposed as a superior approach in the context of machine learning and deep learning models due to its exceptional performance and accuracy in identifying hate speech tweets through the integration of GRU, CNN, and RNN.

2. The suggested GCR-NN model can identify hate speech in 97% of tweets.

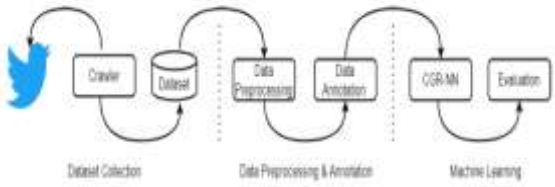

Fig.2: System architecture

Modules:

In order to execute the aforementioned project, we developed the following modules

- Data exploration: we will put data into the system using this module.
- Processing: we will read data for processing using this module.
- Using this module, data will be separated into train and test models.
- Model generation: GCN with BERT, LSTM, GRU, RNN, CNN, Ensemble Method LSTM + GCN with BERT, Logistic Regression, Random Forest, KNN, Decision Tree, Support Vector Machine, Voting Classifier.
- The user signup and login module enables users to create an account and log in to access the platform's features and functionality.
- Prediction: the final predicted value will be presented.

IV. IMPLEMENTATION

Algorithms

GCN: It stands for Graph Convolutional Network, which is a type of neural network that is designed to operate on graph data structures. In contrast to traditional neural networks that operate on vectors or matrices, GCNs can process data represented as graphs, which are a common way of modeling complex relationships between entities. GCNs are widely used in tasks such as node classification, link prediction, and graph clustering, and have shown promising results in various fields such as computer vision, natural language processing, and social network analysis.

BERT: BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained natural language processing (NLP) model developed by Google that uses a deep neural network architecture based on the Transformer architecture. It was designed to pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers. The pre-trained model can then be fine-tuned on specific NLP tasks such as text classification, question answering, and named entity recognition, among others. BERT has achieved state-of-the-art results on several benchmark NLP tasks and is widely used in various NLP applications.

LSTM: LSTM stands for Long Short-Term Memory, which is a type of recurrent neural network (RNN) architecture used in deep learning. LSTM networks are designed to handle the issue of vanishing gradients that arises in traditional RNNs, where the model has difficulty retaining information from previous time steps due to the repeated multiplication of gradient values. LSTM networks

address this problem by introducing a "memory cell" that allows the network to selectively retain or forget information over time. This makes LSTM networks particularly effective for tasks that involve sequential or time-dependent data, such as natural language processing, speech recognition, and time series analysis.

GRU: GRU stands for Gated Recurrent Unit, which is a type of recurrent neural network (RNN) that is used for processing sequential data, such as text or time series data. It was introduced as a simpler and more efficient alternative to the standard RNN architecture. Like other RNNs, GRUs are able to remember information from previous inputs and use that information to make predictions about the current input. However, they are able to do so more efficiently by using a gating mechanism that allows them to selectively update and forget information. This makes them well-suited for tasks such as language modeling, machine translation, and speech recognition.

RNN: RNN stands for Recurrent Neural Network. It is a type of neural network that can process sequential data such as time series data or natural language text by using the concept of recurrence. In an RNN, each node or neuron maintains an internal memory or state, which is updated as it processes each element in the sequence. This allows the network to remember the context of previous elements in the sequence as it processes the current element, and use this context to make predictions or classifications. RNNs are widely used in tasks such as language modeling, speech recognition, and machine translation.

CNN:CNN stands for Convolutional Neural Network, which is a type of deep neural network that is commonly used for image classification, object detection, and natural language processing. The network architecture of CNNs includes convolutional layers that apply filters to input data, pooling layers that reduce the size of feature maps, and fully connected layers that classify the input based on the learned features. CNNs use a hierarchical approach to learning, where lower-level features are learned first and then combined to form higher-level features. This allows CNNs to automatically extract relevant features from input data, making them well-suited for tasks that involve processing large amounts of complex data.

Ensemble Method: Ensemble methods are machine learning techniques that involve combining multiple individual models to improve overall performance. The basic idea behind ensemble methods is that by combining the predictions of multiple models, the strengths of each individual model can be exploited while minimizing their weaknesses, resulting in a more accurate and robust final prediction. Ensemble methods can be used with a wide range of models, including decision trees, neural networks, and support vector machines, among others. Common ensemble methods include bagging, boosting, and stacking, each with its own specific approach to combining models.

Logistic Regression: Logistic regression is a statistical method used to analyze the relationship between a dependent variable (target) and one or more independent variables (predictors) when the dependent variable is binary or dichotomous. It is a type of regression analysis that is used to estimate the probability of a categorical dependent variable. Logistic regression works by using a logistic function to model the relationship between the dependent variable and the independent variables. The logistic function transforms a linear combination of the independent variables into a value between 0 and 1, representing the probability of the dependent variable being in a certain category. It is widely used in fields such as epidemiology, medical research, social sciences, and engineering.

Random Forest: Random forest is a machine learning algorithm that belongs to the family of ensemble learning methods. It is based on building multiple decision trees and combining their results to make a final prediction. Each tree in the forest is constructed using a random subset of the training data and a random subset of the features. The output of the random forest is determined by averaging the output of each individual tree or by taking a majority vote. Random forest is often used for classification and regression tasks, and is known for its ability to handle high-dimensional data and avoid overfitting.

KNN: KNN stands for K-Nearest Neighbors. It is a non-parametric machine learning algorithm used for classification and regression. In this algorithm, the data points are classified based on their proximity to the K nearest data points. The value of K is a hyperparameter that is predefined by the user. KNN algorithm is based on the assumption that data points that are close to each other are likely to belong to the same class or have similar characteristics.

Decision tree: A decision tree is a non-parametric supervised learning algorithm used for classification and regression analysis. It is a tree-structured model where internal nodes represent the attributes/features of the dataset, and the branches represent the corresponding value of these attributes. The leaves of the tree represent the decision outcomes, and each path from the root of the tree to a leaf represents a decision rule. The goal of the decision tree algorithm is to create a model that predicts the target variable by learning simple decision rules from the data features. The decision tree algorithm can handle both categorical and numerical data and is simple to understand and interpret, making it a popular choice for many machine learning tasks.

SVM: SVM (Support Vector Machine) is a supervised machine learning algorithm that can be used for classification or regression tasks. In SVM, a set of training data is used to train the model, and then it can be used to predict the classes or values for new data. The algorithm works by finding the optimal hyperplane that separates the data into different classes, with the goal of maximizing the margin between the classes. SVM can handle both linear and non-linear data by using kernel functions to map the data into a higher dimensional space. SVM is often used for tasks such as image classification, text classification, and bioinformatics analysis.

Voting classifier: A voting classifier is an ensemble method in machine learning where multiple models (e.g., logistic regression, decision tree, random forest, etc.) are used together to make predictions. Each model produces its own prediction based on the input data, and then the voting classifier combines the predictions of all models and selects the most frequent class as the final prediction. Voting classifiers can improve the accuracy and reliability of predictions by leveraging the strengths of multiple models and compensating for the weaknesses of individual models.

They are commonly used in classification tasks where a single model may not be sufficient to accurately classify the data.

V. PROJECT-RESULTS

Fig.3: Home screen

Fig.4: User registration

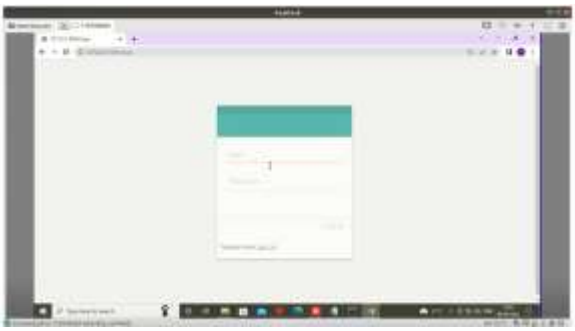

Fig.5: user login

Fig.6: Main screen

Fig.7: User input

Fig.8: Prediction result

VI. CONCLUSION

Hate speech has become increasingly prevalent on social media platforms like Twitter, making it critical to automatically detect and block such content to prevent its further spread. To achieve this, sentiment analysis is used to identify negative sentiments and hate tweets. The proposed solution leverages the ensemble technique, which stacks GRU, CNN, and RNN to build the GCR-NN model, resulting in high-performance sentiment analysis. A large dataset of 169,999 tweets collected from Twitter and annotated with TextBlob was used to evaluate various machine learning, deep learning, and GCR-NN models. The study found hate speech in 31.49% of the tweets analyzed. The results showed that deep learning models outperformed machine learning models, with the suggested GCR-NN model achieving an average accuracy score of 0.98 for positive, negative, and neutral sentiment classifications. Since the negative class is critical in detecting hate speech, a secondary analysis was performed to evaluate SVM and LR models, which correctly identified 96% and 95% of hate tweets, respectively, but misclassified 4% and 5% of hate tweets. In contrast, the suggested GCR-NN model accurately detected 97% of hate tweets with a minimal 3% error rate.The prevalence of hate speech on social media platforms, particularly Twitter, is a growing concern that has led to efforts to develop automatic detection and blocking mechanisms. This study presents a novel approach to detecting hate speech, using sentiment analysis to identify negative tweets that contain discriminatory language. To achieve high accuracy in this task, the study employs deep learning models and the ensemble technique to build a Gated Convolutional Recurrent-Neural Network (GCR-NN) that combines the strengths of gated recurrent units (GRU), convolutional neural networks (CNN), and recurrent neural networks (RNN).

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of 1] K. R. Kaiser, D. M. Kaiser, R. M. Kaiser, and A. M. Rackham, ‘‘Using social media to understand and guide the treatment of racist ideology,’’ Global J. Guid. Counseling Schools, Current Perspect., vol. 8, no. 1, pp. 38–49, Apr. 2018.

[2] A. Perrin and M. Anderson. (2018). Share of U.S. Adults Using Social Media, Including Facebook, is Mostly Unchanged Since 2018. [Online]. Available: https://www.pewresearch.org/fact-tank/2019/04/10/share-ofu-s-adults-using-social-media-including-facebook-is-mostly-unchangedsince-2018/

[3] M. Ahlgren. 40C Twitter Statistics & Facts. Accessed: Sep. 1, 2021. [Online]. Available: https://www.websitehostingrating.com/twitterstatistics/

[4] D. Arigo, S. Pagoto, L. Carter-Harris, S. E. Lillie, and C. Nebeker, ‘‘Using social media for health research: Methodological and ethical considerations for recruitment and intervention delivery,’’ Digit. Health, vol. 4, Jan. 2018, Art. no. 205520761877175.

[5] Pazhani. A, A. J., Gunasekaran, P., Shanmuganathan, V., Lim, S., Madasamy, K., Manoharan, R., &Verma, A. (2022).Peer–Peer Communication Using Novel Slice Handover Algorithm for 5G Wireless Networks.Journal of Sensor and Actuator Networks, 11(4), 82.

[6] M. A. Price, J. R. Weisz, S. McKetta, N. L. Hollinsaid, M. R. Lattanner, A. E. Reid, and M. L. Hatzenbuehler, ‘‘Meta-analysis: Are psychotherapies less effective for black youth in communities with higher levels of anti-black racism?’’ J. Amer. Acad. Child Adolescent Psychiatry, 2021, doi: 10.1016/j.jaac.2021.07.808. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0890856721012818

[7] D. Williams and L. Cooper, ‘‘Reducing racial inequities in health: Using what we already know to take action,’’ Int. J. Environ. Res. Public Health, vol. 16, no. 4, p. 606, Feb. 2019.

[8] Y. Paradies, J. Ben, N. Denson, A. Elias, N. Priest, A. Pieterse, A. Gupta, M. Kelaher, and G. Gee, ‘‘Racism as a determinant of health: A systematic review and meta-analysis,’’ PLoS ONE, vol. 10, no. 9, Sep. 2015, Art. no. e0138511.

[9] J. C. Phelan and B. G. Link, ‘‘Is racism a fundamental cause of inequalities in health?’’ Annu. Rev. Sociol., vol. 41, no. 1, pp. 311–330, Aug. 2015.

[10] D. R. Williams, ‘‘Race and health: Basic questions, emerging directions,’’ Ann. Epidemiol., vol. 7, no. 5, pp. 322–333, Jul. 1997.

[11] Z. D. Bailey, N. Krieger, M. Agénor, J. Graves, N. Linos, and M. T. Bassett, ‘‘Structural racism and health inequities in the USA: Evidence and interventions,’’ Lancet, vol. 389, no. 10077, pp. 1453–1463, Apr. 2017.

[12] D. R. Williams, J. A. Lawrence, B. A. Davis, and C. Vu, ‘‘Understanding how discrimination can affect health,’’ Health Services Res., vol. 54, no. S2, pp. 1374–1388, Dec. 2019.

[13] C. P. Jones, ‘‘Levels of racism: A theoretic framework and a gardener’s tale,’’ Amer. J. Public Health, vol. 90, no. 8, p. 1212, 2000.

[14] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.

[15] B. J. Goosby, J. E. Cheadle, and C. Mitchell, ‘‘Stress-related biosocial mechanisms of discrimination and African American health inequities,’’ Annu. Rev. Sociol., vol. 44, no. 1, pp. 319–340, Jul. 2018.