International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

SENTIMENT BASED RACISM DETECTION OF TWEETS USING DEEP LEARNING MODULES

ABDUL NAZMA

Master of Computer Applications (MCA),

SVKP & Dr.K.S Raju Arts & Science College(A),

Penugonda, W.G.Dt., A.P, India

abdulnazma999@gmail.com

CH SRINIVASA RAO

Associate Professor in Computer Science,

SVKP & Dr.K.S Raju Arts & Science College(A),

Penugonda, W.G.Dt., A.P, India

chiraparapu@gmail.com

ABSTRACT:

Due to social media's dominant position in the sociopolitical environment, a number of traditional and contemporary types of racism were practised there. Racism has surfaced on social media in a variety of hidden and open ways, including the use of memes to hide it and racist remarks that exploit fictitious identities to stir up violence, hatred, and social unrest. Racism now thrives on the basis of colour, origin, language, cultures, and most crucially religion, despite the fact that it is frequently associated with ethnicity. Social media thoughts and comments that incite racial animosity have been seen as posing a severe threat to social, political, and cultural stability as well as the peace in various nations. Therefore, as social media is the main platform for the transmission of racist ideas, it should be closely watched, and any racist remarks should be quickly identified and blocked. By using sentiment analysis on Twitter, this study aims to pinpoint racist tweets. Due to deep learning's improved performance, gated recurrent units (GRU), convolutional neural networks (CNN), and recurrent neural networks (RNN) are combined to create gated convolutional recurrent- neural networks (GCR-NN), which are a stacked ensemble deep learning model. In this GCR-NN model, GRU performs best at extracting relevant and salient characteristics from unprocessed text, whereas CNN extracts critical features for RNN to generate precise predictions. Therefore, a number of experiments are carried out to examine and evaluate the effectiveness of the suggested GCR-NN. Numerous tests have been carried out to examine and evaluate the outcome of the proposed GCR-NN within the context of machine learning and deep learning models, and the results clearly show that the GCR-NN performs better than both of these models with an improved 0.98 accuracy. The suggested GCR-NN model is capable of identifying 97% of tweets with racist content.

KEYWORDS: Racism, Social media, online abuse, Twitter, Deep Learning.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

I.INTRODUCTION

Social media has become a dominating element in socio-political prospects and controls our minds and actions in different ways. With the wide use of social media platforms over the world and freedom of speech, several vices have emerged over the past few years, racism being one of the leading ones. Social media sites, such as Twitter, represent a new setting in which racism and related stress are apparently prospering [1]. Currently, 22% of United States (US) adults use Twitter [2], while Twitter has 1.3 billion accounts and 336 million active users across the globe, 90% of which has a public profile leading to 500 million tweets per day [3]. Unless tweets are made private, they are publicly available and Twitter users can react to such tweets and engage by sharing them on their profile (re tweet), tagging someone's user name, clicking the like button, or responding to the author of the tweet [4]. In Twitter, the expression of feelings, emotions, attitudes, and opinions build the raw data of sentimental analysis [5].

The widespread usage of social media platforms for many contemporary and historical kinds of racism is a result of their rising popularity [6]. Racism is represented on these platforms in a variety of covert ways, like through memes, and outwardly, as by publishing racist Tweets under false names. Racism now thrives on the basis of color, origin, language, cultures, and most crucially religion, despite the fact that it is frequently associated with ethnicity. Social media thoughts and comments that incite racial animosity have been seen as posing a severe threat to social, political, and cultural stability as well as the peace in various nations. Since social media is the main platform for the transmission of racist ideas, it should be closely watched, and any racist remarks should be quickly identified and blocked.

Racist comments and tweets on social media have been regarded as the source of several kinds of mental and body illness leading to adverse health outcomes [7]_[12]. With respect to its use on social media, racism can be categorized into three groups: institutionalized, personally mediated, and internalized [13]. Personally mediated racism can be experienced through racial discrimination or differential racial treatment, or through awareness of discrimination against family and friends. Consequently, the racist behavior of the society adversely affects individuals and ignites several kinds of psycho-social stress often leading to the risk of chronic diseases [14][16]. Additionally, racist groups and individuals perpetuate cyber-racism by employing higher skill levels and

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

intricacy through various channels and strategies [5].Special considerations have been given to the field of sentiment analysis to analyze the text from social media platforms for a large variety of tasks including hatred speech detection, market prediction based on sentiments, and racism detection, etc.

The wide use of social media is a potential source of data generation containing important information regarding people's attitudes, responses, emotions, and opinions regarding specific events, objects, personalities, and entities. Sentiment analysis provides powerful tools to mine such data to analyze emotions. The huge part of Twitter feeds become less characterized by coherent rational discussion, but more by foods of emotion and affect, and can be used to divide the narratives into polarities of good and evil [17], [18]. Research shows that issues may become less obvious than a shared sense of outrage and a compelling sense of shared agreement and Twitter feeds can be quite insular and nodal [19].

Keeping in view its wide use, social media has become an attractive source to apprehend attitudes and analyze interactions over sensitive topics such as racism. In the USA, the discussions about race and ethnicity on Twitter have been considered as indicators of the current state of relations based on race. Additionally, the variation in the types of discussions about racism indicates the geographic variability in racial attitudes and sentiment [20]. So, analyzing the details of how people, events, and circumstances are represented reveals the dynamics of how users communicate, and many problems related to racism can be exposed on this platform.

Owing to the extreme and atypical racist attitude an individual faces related to personal traits and attitudes, one can easily become relativized, contextualized, and therefore depoliticized. It leads to distracting attention from the actual and specific structural inequalities in society experienced by certain ethnic groups [21].

Machine and deep learning approaches has proven their strength and superiority over traditional methods in several domains such as image processing [22], [23], text classification [24], [25] and sentiment analysis is no exception. Several recent studies show that machine learning techniques perform better for sentiment analysis tasks [26], [27]. Therefore, this study leverage machine learning and deep learning models to perform sentiment analysis on tweets related to racism and makes the following contributions

_ An ensemble model is proposed that makes use of recurrent neural networks. For this purpose, gated recurrent unit (GRU), convolution neural network, and recurrent neural

network are stacked to make the GCR-NN model to perform sentiment analysis.

_ A large dataset of tweets containing racist comments/ text is crawled from Twitter which can be used by the research community. The dataset is annotated using the Text Blob based on the polarity score into positive, negative, and neutral sentiments.

_ For performance comparison, several well-known machine learning models are implemented using the optimized parameters such as decision tree (DT), random forest (RF), logistic regression (LR), k nearest neighbor (KNN), and support vector machines (SVM). Term frequency-inverse document frequency (TF-IDF) and bag of words (BOW) are studied as feature extraction techniques.

_ For a fair comparison with the proposed approach, GRU, long short term memory (LSTM), CNN, and RNN are implemented as standalone models. Similarly, the performance of several state-of-the-art models is compared with the proposed GCR-NN in terms of accuracy, precision, recall, and F1 score.

II. LITERATURE SURVEY

The overwhelming effects of hate crimes are increasing to a great extent because of the extensive use of social media [37] and the anonymity enjoyed by online users [38]. Abusive content and intricate stuffing on social media is a problematic phenomenon with more than a few overlapping and coinciding modes and aims [31]. The contents related to harassment and maltreatment arouses negative feelings in online users so they express their feelings in a discourteous way. Cyberbullying and hate speeches are two examples of abusive languages that have vexed the interest of researchers in recent times owing to their harmful effects on society. Decontamination of these contents is very necessary. For this purpose, several studies have been conducted to automatically detect the annoying hate speeches and massages among other contents on social media. Automatic hate speech detection using machine learning algorithms is still new and requires extensive research efforts from both industry and academia [39]. Few recent and related papers have been discussed here [40], [41]. Machine learning algorithms have contributed enormously to hate speech detection and content analysis [37].

A multimodal hate speech detection algorithm tailored for Greek social media is presented by the authors in [28]. The study focuses on Greek-language tweets that criticize immigrants and refugees, particularly those that employ racist and xenophobic language. On the gathered dataset, the ensemble model, transfer learning, and fine-tuning of the BERT and Resnet bidirectional encoder representations are used.

The highest accuracy was reported with nlpaueb/greek-bert for text modality and 0.97 with resnet18+ nlpaueb/greek-bert for text+image modality. Different variations of the BERT and Resnet are employed. A similar cutting-edge machine learning-based solution for the automatic detection of hate speech in Arabic social media networks is suggested by [29]. Different sets of features are employed for analysis while various emotional states are recorded. The study employs Naive Bayes (NB), DT, SVM, and RF with TF-IDF, profile-related, and emotion-related data, as well as four distinct machine learning techniques. The maximum accuracy was attained by RF using TF-IDF and profile-related features, which was 0.913.

Along the same lines, [30] classifies the fake news and hate speech propaganda using the extracted features from the content containing fake and real news. The study uses NB, LR, and XGBoost with TF-IDF features. XGBoost demonstrates a recall value of 0.83 which indicates that 17% of data contains hatred content and is misclassified by the model. Also, XGBoost achieves the precision value of 0.82 which shows that 18% of data is hateful and the model misclassified it. Authors investigate the hate speech problem in the Saudi Twitter sphere in [31] using different deep learning approaches. A series of experiments are conducted on two datasets using BERT, CNN, GRU, and the ensemble of CNN and GRU (CNN+GRU). Results indicate that the model achieves an F1 score of 0.79 and the area under receiver operating curve (AUROC) of 0.89 using the CNN model.

Study [32] investigates the automatic detection of cyberbullying. To review the deep learning and machine learning approaches, the authors use two different datasets. Different word embedding techniques such as distributed BoW (DBoW), distributed memory mean (DMM) and Word2Vec CNN are used to classify online racism. An accuracy of 96.67% for one dataset while 97.5% for the second dataset is achieved using a neural network with 3 hidden layers using Doc2Vec features. In the same way, study [33] explores the automatic detection of Indonesian tweets that contain hate speech or racism. The authors use machine learning models such as multinomial NB (MNB), Multilayer Perceptron (MLP), AdaBoost (AB) classifier, and SVM. Synthetic minority oversampling technique (SMOTE) is used as an upsampling technique and experiments are performed on both SMOTE and non-SMOTE features. Results show that MLP with SMOTE features has an accuracy of 83.4% and AB, and MNB has 71.2% accuracy for non-SMOTE features.

Ching She et al. work on hate speech detection from social media in [34]. For experiments, the

audio data is extracted from videos and converted to text using a speech-to-text converter. MNB, Linear SVM, RF, and RNN are used for experiments. Two different sets of experiments are carried out where the first experiment involves classifying the video into normal and hateful videos while the second experiment aims at classifying the video into normal, racist, and sexiest classes.

Results show that RF shows superior performance in terms of accuracy and achieves an accuracy of 0.9464 for the first set of experiments and 0.857 for the second set of experiments.

Another similar work is [35] which investigates hate speech related to Islam on social media. The study constructs an automated tool that can distinguish between non islamophobic, weak islamophobic, and strong islamophobic content. Different machine learning algorithms such as NB, RF, LR, DT, SVM, and deep learning models are used. Results suggest that SVM obtains the testing accuracy of 72.17%. The performance of SVM is also evaluated using 10 fold cross-validation which shows a 74.6% accuracy and balanced accuracy of 80.7%. Study [36] proposes a novel system to detect hate speech across multiple social media plate forms like Reddit, YouTube, Twitter, and Wikipedia. A large dataset is built from these social media platforms with 80% labeled as non-hateful and 20% labeled as hateful. Several machine learning algorithms such as XGBoost, SVM, LR, NB, and feed-forward neural networks and tested with BoW, TF-IDF, Word2Vec, BERT, and their combinations. XGBoost outperforms all models with a 0.92 F1 score with all features. Feature importance analysis shows that BERT features have a great effect on predictions.

Taking into account the reported results from deep learning models, this study leverages the deep learning ensemble model to detect racism comments from Twitter. The study aims at obtaining high classification accuracy by stacking recurrent neural networks. Racism detection is performed using sentiment analysis where the ratio of tweets containing negative sentiments indicates the racist tweets.

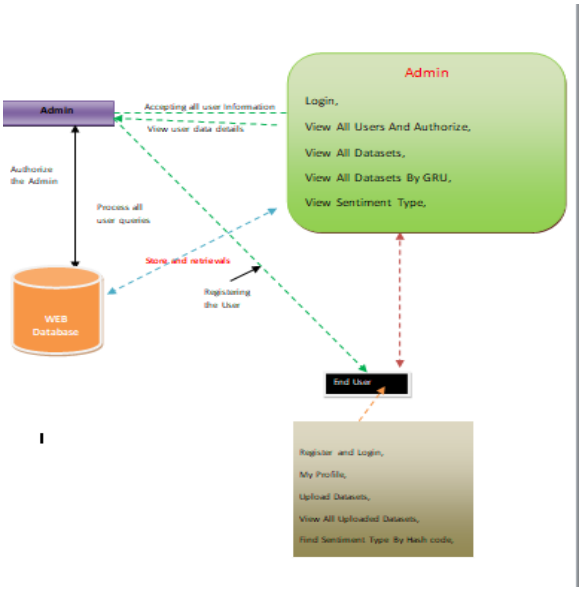
III.PROBLEM STATEMENT

A multimodal hate speech detection algorithm tailored for Greek social media is presented by the authors in [28]. The study focuses on Greek-language tweets that criticize immigrants and refugees, particularly those that employ racist and xenophobic language. On the gathered dataset, the ensemble model, transfer learning, and fine-tuning of the BERT and Resnet bidirectional encoder representations are used. The best accuracy of 0.944 is recorded with nlpaueb/greek-bert for the text modality and 0.97 with resnet18C nlpaueb/greek-bert for the text Cimage modality utilizing several variants

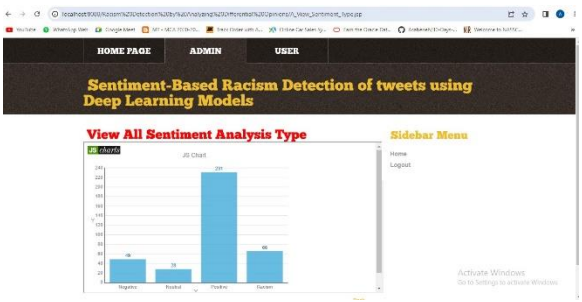
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

of the BERT and Resnet. A similar cutting-edge machine learning-based solution for the automatic detection of hate speech in Arabic social media networks is suggested by [29]. Different sets of features are employed for analysis while various emotional states are recorded. The study employs Naive Bayes (NB), DT, SVM, and RF with TF-IDF, profile-related, and emotion-related data, as well as four distinct machine learning techniques. The most accurate method for detecting hate speech in Arabic social media networks, with a detection accuracy of 0.913.atic, was RF with TF-IDF and profile-related characteristics. Along the same lines, [30] classifies the fake news and hate speech propaganda using the extracted features from the content containing fake and real news. The study uses NB, LR, and XGBoost with TF-IDF features. XGBoost demonstrates a recall value of 0.83 which indicates that 17% of data contains hatred content and is misclassified by the model. Also, XGBoost achieves the precision value of 0.82 which shows that 18% of data is hateful and the model misclassified it. Authors investigate the hate speech problem in the Saudi Twitter sphere in [31] using different deep learning approaches. A series of experiments are conducted on two datasets using BERT, CNN, GRU, and the ensemble of CNN and GRU (CNNCGRU). Results indicate that the model achieves an F1 score of 0.79 and the area under receiver operating curve (AUROC) of 0.89 using the CNN model.

ARCHITECTURE:



IV.RESULTS



V.CONCLUSION

Racist remarks are more common on social media sites like Twitter and should be automatically identified and blocked in order to stop them from spreading. In this study, racism is detected using sentiment analysis to identify tweets that include racist content by identifying unfavourable feelings. Deep learning is augmented by the ensemble method, in which GRU, CNN, and RNN are layered to create the GCR-NN model, in order to provide high-performance sentiment analysis. For experimentation with various machine learning, deep learning, and the proposed GCR-NN model, a sizable dataset gathered from Twitter and annotated using Text Blob is employed. In total, 169,999 tweets were collected, and 31.49% of them contained racial remarks. The suggested GCR-NN obtained an average accuracy score of 0.98 regarding the sentiment analysis for positive, negative, and neutral classes, demonstrating that deep learning models perform significantly better than those of machine learning models. Given that the negative class is crucial for detecting racism, a separate research shows that SVM and LR are able to accurately identify 96% and 95% of racist tweets, respectively, whereas 4% and 5% of racist tweets are misclassified. On the other hand, the suggested GCR-NN can accurately identify 97% of the racist tweets with just a 3% misclassification rate.

REFERENCES

[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] A.-M. Bliuc, N. Faulkner, A. Jakubowicz, and C. McGarty, ‘‘Online networks of racial hate: A systematic review of 10 years of research on cyberracism,’’ Comput. Hum. Behav., vol. 87, pp. 75–86, Oct. 2018.

[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] S. Forrester, D. Jacobs, R. Zmora, P. Schreiner, V. Roger, and C. I. Kiefe, ‘‘Racial differences in weathering and its associations with psychosocial stress: The CARDIA study,’’ SSM-Population Health, vol. 7, Apr. 2019, Art. no. 100319.

[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.

[16] A. T. Geronimus, M. Hicken, D. Keene, and J. Bound, ‘‘‘Weathering’ and age patterns of

allostatic load scores among blacks and whites in the United States,'' Amer. J. Public Health, vol. 96, no. 5, pp. 826–833, 2006.

[17] Z. Papacharissi, ''Affective publics and structures of storytelling: Sentiment, events and mediality,'' Inf., Commun. Soc., vol. 19, no. 3, pp. 307–324, Mar. 2016.

[18] G. Bouvier, ''How journalists source trending social media feeds: A critical discourse perspective on Twitter,'' Journalism Stud., vol. 20, no. 2, pp. 212–231, Jan. 2019.

ABOUT AUTHORS:

ABDUL NAZMA currently pursuing MCA in SVKP and Dr.K.S.Raju Arts & Science College(Autonomous) affiliated to Adikavi Nannaya University, Rajamahendravaram. Her research interests includes Data Structures,DataMining,WebTechnologies and Artificial Inteligence.

CH.SRINIVASA RAO is working as Associate Professor in SVKP & Dr K S Raju Arts & Science College(A), Penugonda , West Godavari District, A.P. He received Master's Degree in Computer Science & Engineering(M.Tech, CSE) from Jawaharlal Nehru Technological University, Kakinada, India. He qualified in UGC NET and AP SET. His research interests include DataMining, Artificial Inteligence and WebTechonologies.