iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81277

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Ethical AI Analyzer

Nityam Kumar Tiwari, Neeraj Patel, Himanshu, Nistha Garg, Mr. Jagbeer Singh
Department of Computer Science and Engineering Meerut Institute Of Engineering and Technology, Meerut, Uttar Pradesh, India

Abstract: ***Human biases significantly influence the efficiency and fairness of Natural Language Processing(NLP)systems.Socialbiases,encompassinggender,race, and caste frequently manifest within linguistic datasets and are subsequently potentiallyyieldinginequitableresults.Althoughtheinvestigationofbias within textual data has expended in recent years, existing resources remain constrained, fragmented and predominantly concentrated on specific categories such as gender or race with insufficient consideration given to caste or the interplay of multiplebiases.***
This study has two main goals: first, to investigate how large languagemodels (LLMs) and transformer-based architectures can be used to find and categories different types of social bias in text.
Second,toanalyzethechallengesof limited datasets, uneven distribution of categories, and the difficulties of fair evaluation. The results highlights the need for more diverseandrepresentativedatasets,andtheyprovideinsightsintocreatingmorefairand inclusive Natural Language Processing.

I. INTRODUCTION

There is growing recognition that human biases are deeply embedded in languageandcanstrongly affectthebehaviorandfairnessofNaturalLanguage Processing(NLP)systems.Thesebiasesoftenreflectbroadersocialinequalities and when learned by machine learning models, can result in discriminatory outcomes in real-world application. As a result, research onbias detection, classification, and mitigation in NLP has gained increasing attention, with broader goal of developing more fair and responsible AI systems.

Despitethisprogress,biasdetectioninNLPremainsinsufficiently explored.Many existing datasets are smallin scale, focus on a limited set of categoriessuch asgenderorrace,or wereoriginally designedforrelated tasks like hate speech detectionrathert hanexplicitbiasidentification.

Context-dependent forms of bias, particularly caste-based bias, are largelyabsent from mainstream resources.

These gap raise an important question: How can NLP models be effectively trainedto detect and classify multiple formsofsocial biasusing limited and imperfect data?

To address this question, this study pursues three main objectives. First, it analyzespublicly availabledatasetsrelatedtobiasandhatespeechand usesthemas trainingresources.Second,itintroducesacustom-annotated datasets centered on caste bias, extending the scope beyond commonly studied categories. Third, trains and evaluates transformer-based model and large language model (LLMs) on both binary (biased vs. non-biased) and mulit-class (e,g. Gender, race, caste) classification tasks.

Thisworkcontributestoexistingresearchbyprovidingacomparativeanalysisof model performance across different bias categories, highlighting the impactof datasets composition and class imbalance. It also discuss the challenges associatedwith,heterogeneous,andunderrepresentativedata,emphasizingthe need for more balanced and intersectional datasets.

In this study, bias is defined as unequal or unfair treatment based on social characteristics. We consider bias expressed through derogatory language, harmful stereotypes, and abusive or exclusionary statements targetingspecific groups. The target categories examined include gender, race, caste, religion, disability,sexualorientation,nationallyandage.Whilebiasandhatespeechare closelyrelated,theyarenotequivalent;therefore,resourcesfrombothdomains are utilized”

We acknowledge the sensitive nature of studying social bias and recognize existing limitations, such as simplified representations of gender and caste and limited modeling of intersectionality due to dataset constraints. The remainder of the paper is organized as follows : Section 2 reviews related work,section3describesdatasetconstructionandpreprocessing,Section4 details the modeling and evaluation framework, section 5 discuss the result, and section 6 concludes with directions for future research.

II. RELATED WORK

We begin our work by reviewing existing research on bias and hate speech detection in Natural Language Processing (NLP), with particularly attentioncommonly used models, datasets, and their limitations.

©IJRASET: All Rights are Reserved

We also examine real-worldexamplesofbiaslanguagedrawnfromonlineplateformssuchas Twitter(X)and LinkedIn which illustrate how social bias appears in everyday digital communication.

We observe that research on bias in NLP has expanded rapidly in recent years, especially with the rise of transformer-based architecture such as BERT and its variants. Prior studies have demonstrated that these models can encoded and propagate harmful stereotypes present in their training data. Bias has been reported across a wide range of NLP task, including coreference resolution, sentiment analysis, dialogue generation, and syntheticparsing.Thesefindingssuggestthatbiasisnotconfinedtospecific applications but is instead a broader challenge affecting modern NLP systems as a whole.

We identify training data as a major contributing factor to this prpblem. Large-scale datasets collected from web crawls and social media plateforms contain rich and diverse language, but they also include subsequential amounts of offensiveanddiscriminatorycontent.Forinstances,analysesof widelyused corporate such as Common Crawlhave shown that a noticeable portion of the data contains hate speech or sexually explicit material. Models trained om such data may therefore unintentionally learn baised associations. High-profile incidents involving systems such as Microsoft's Tay chatbot,Meta's Galactica, and other large language models further highlight the risk of deploying biased AI systems, even when mitigation strategies are applied.

Atthesametime,wenoteagrowingbodyofworkthatfocusesonusing NLP models themselves to detect and classify bias in legal tetxs, study the representation of marginalized communities in Wikipedia, and examine disparitiesin literary reviews across different social groups. Much of this research relies on data collected from social media platform, particularly Twitter, using keyword- based collection techniques. More recent studies haveexploredsyntheticdatagenerationtoaddressdatascarcity, although this data generation to addresses data scarcity, although this approach raises concerns about realism and representativeness.

Overall, we find that existing datasets are often limited in size, narrowly focusedonspecificbiascategories,ororiginallydesignedforrelated tasks such as hate speech detection rather than explicit bias classification. Theselimitationsmotivateourworkandunderscoretheneedforbroader,more representative datasets and systematic evaluations across multiple forms of social bias.

III. METHODOLOGY

In this section, we describe the design and implementation of our system for detectingsocialbiasinonlinetextualcontent.Ourmethodologyisorganizedinto four main stages: (1) data collection and preprocessing, (2) definition of target bias categories, (3) model selection and training, and (4) inference and output generation.

A. Problem Definition

Theobjectiveofthisworkistoautomaticallyidentifybiasedlanguageinusergeneratedonlinecontent,withparticularemphasisoncaste-based, genderbased,religious,andracialbias.Suchformsofdiscriminationfrequentlyappearinsocialmediapostsandprofessionaldiscussions,eithe rexplicitlyorinmore subtle forms. Unlike conventional hate speech detection tasks, ourapproachesaimstocapture bothoverthostilityandimplicitbiasedframing, evenwhenh armfulinten tisexpressedindirectly.

B. DataCollectionandprocessing

Wecollectedtextualdatafrompubliclyavailablesources,includingplatforms such as Twitter and LinkedIn, where biased language is commonly observed. In addition, we validation and analysis. To ensures ethical compliances and protectuserprivacy,allidentifiableinformationwasremovedor anonymized.

Thecollectedtextwascleanedandnormalizedusingstandardpreprocessing steps, including lowercasing, removal of punctuation, URLs, hashtags, mentions, and emojis, and tokenization using a transformer-basedtokenizer with sub-word segmentation. The resulting corpus combines curated benchmark data with real- world examples of both biased and non- biased language, enabling robust and generalizable model training.

C. TargetBiasCategories

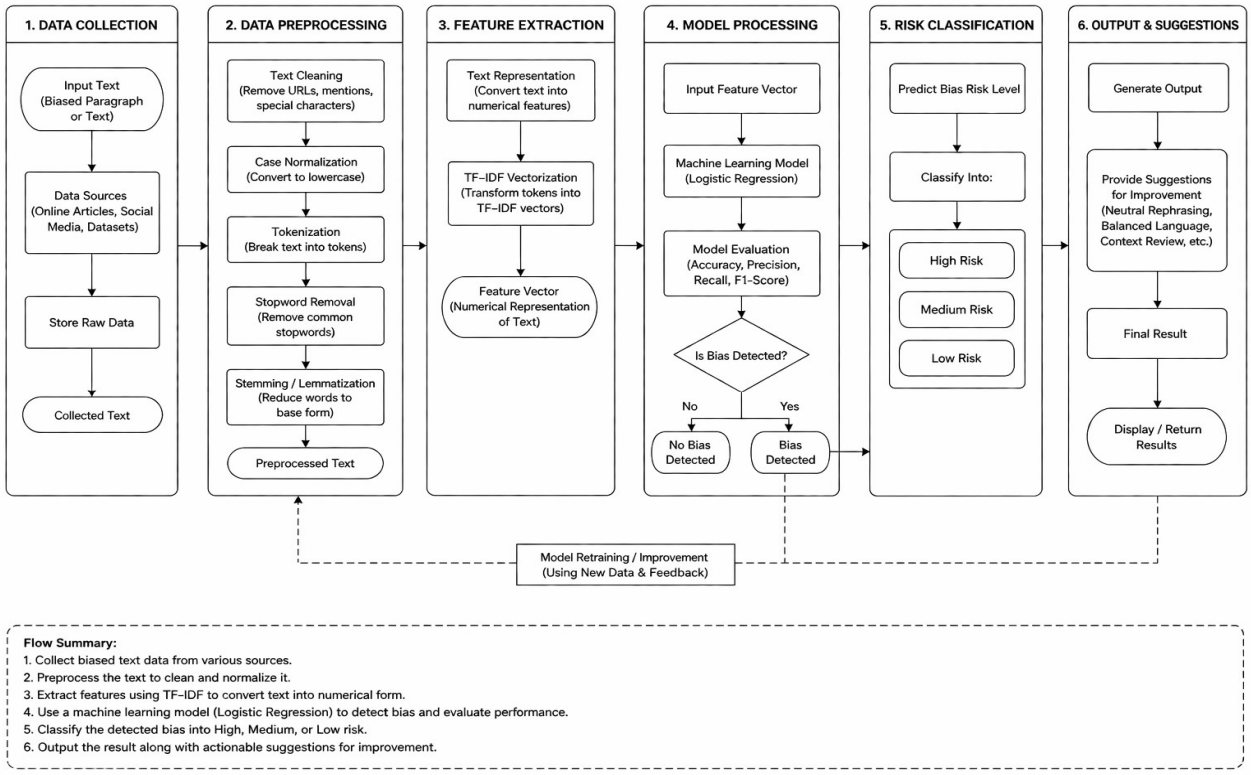
Wecategoriesbiasbasedonthesocialgroupbeingtargeted.Theprimarytarget categories considered in this study include gender, caste, race and ethnicity, religion, and profession-related bias. Each sample was annotatedwith a binary labelindicatingwhetheritisbiasedornon-biased.When sufficient contextual or linguistic cues were available, the system further classified the specific typesofbias.

D. ModelArchitecture

For model implementation, we employed an Emotion-Transformer architecture built on top of Distill BERT, selected for its balance between computational efficiency and strong contextual representation. The model was fine-tuned on our dataset for text classification. The final hidden layer outputispassedthroughapoolingoperationfollowedbyaclassificationhead, producing a binary prediction (biased vs. non-biased)and, when applicable, anassociatebiascategory.

E. Flowchart

We illustrate the overall workflow of our system through this flowchart. We begin by taking textual input from the user and initially checkingwhetherthecontentisbiasedorunbiased.Ifthetextisfoundtobeunbiased,weclassifyitaslowornoriskandacceptisasvalid.Whenbias edcontentisdetected,wefurther process the text using NLP techniques to analyze its severity. We then assess thelevelofbiasandcategorizeitintodifferentrisklevels.Forhigh-riskcases,we suggest an alternatives unbiased version of the text, while low and medium risk casesarehandledaccordingly.Thisflowallowsustobothdetectbiasedlanguage and support bias mitigation in a structured manner.



IV. MODEL TRAINING

A. Experimental Setup

Todevelopour biasdetectionsystem, weadoptedtheEmotion-Transfer model , which was originally proposed for emotion classification but can be effectively adapted to indentify biased language in real-world online content, including tweets, LinkdIn posts,and otherpublicly available social media text. Theachitectureisbuilton DistillBERT,adistilledvariantofBERTthatoffersa fevourable balance between computational efficiency and contextual representation capability.

Weimplementedourapproachusingtheopen-sourceEmotion-Transformer framework and the corresponding Distill BERT model provided by the HuggingFaceTransformerslibrary.Toassessthemodel'seffectiveness onestablishedbenchmarks, wetrainedand evaluateditonsubsetsofthe DavidsonandMLMAdatasets.Thesedatasetswereselectedbecausetheir original studies reported clear performance baselines, such as Waseem-Hovy and DynGen,werenot include in thiscomparisondue to differencein task formulation and annotation schemes.

On the Davidson dataset, the original work reported an F1-score of 0.90 usingasupportvector machinewithL2regularization.Usingourtransformer-based configuration,trainedforfiveachievedanF1-scoreof0.80.Similarly, for the MLMA dataset, which originally reported an F1-score of 0.43, our model achievedacomparable F1-scoreof0.42whentrained forfourepochsunder thesameconfiguration.TheseresultsindicatethattheEmotion-Transformer performscompetitivelywhich existingapproachesandissuitableforthe deployment in real-world bias detection scenarios.

V. RESULT AND DISCUSSION

In this section we, present the results obtained from our trained model and discuss the key observations that emerged during our experiments. As a team, we aimed not only to evaluate how accurately the model classifies biased content but also to understand how well it generalizes to different forms ofsocialbiasencounteredonreal-worldplatformssuchasTwitterandLinkedIn.

A. PerformanceonBenchmarksDatasets

Toestablishareliableperformancebaseline,weevaluatedourmodelontwo publicly available benchmark datasets: Davidson and MLMA. We selected thesedatasetsbecausetheiroriginalstudiesreportedF1-scores,allowingus to make direct comparisons. On the Davidson dataset, we achieved an F1-score of 0.80, compared to the originally reported score of 0.90usinganSVMwithL2regularization.OntheMLMAdataset,ourbestresult wasanF1-score of0.42,whichcloselymatchestheoriginal scoreof0.43.Asa team, we interpret these results as evidence that the Emotion-Transformer performs at a level comparable to existing methods on standard hate speech andtoxicitydatasets,v alidatingitsuse forfurtherbiasdetectiontasks.

B. BiasDetectioninReal-WorldText

We consider the application of our model to real-world data from Twitter and LinkedIn to be the most significant contribution of our work. While analyzingthe outputs together,we observedthat these platformsoften containimplicit orcontext – dependent bias that does not rely on overtly offensive language. Afterfine-tuningthemodelusingbothbenchmarkdataandcuratedreal-worldexamples,wefoundthatitconsistentlydetectedbiasedstatementsacross severalcategories. We observed that gender bias was identified with relatively high precision, largely due to the availability of labeled examples. In contrast, wefoundcaste-basedbiastobemorechallenging,especiallywhenexpressed indirectlythroughsubtlelanguageormicroaggressions.However,wenotedthat including manually curated caste-related examples noticeably improved the model'ssensitivity.Wealsoobservedthatreligion-andrace-basedbiaseswere easier to detect when explicit terms were present, whereas amiguousor context-heavycasesremaineddifficult.Overall,weagreedthatthemodelwas more confident when handling highly polarized or toxic statements, while nuanced or sarcastic expression-common in professional setting such as LinkdIn were harder to classify reliably.

C. ChallengesObserved

Throughout our experiments, we collectively identified several challenges that influenced model performance. Class imbalance across bias categories, particularlyforcasteanddisability,limitedthemodel'sabilitytogeneralize. Wealsoencounteredannotationnoiseinreal-worlddata,wherebiaswashighly dependent on context and interpretation. Additionally , we observedthat code mixed language, especially Hindi-English content commonly found on Indian social media platforms, posed difficulties for models primarilytrainedonEnglish text. Despite these challenges , we found that the system was still able to flag a substantial portion of biased content accurately and consistently, reinforcing its potentialvalueforbiasanalysis,awareness,andmoderation-relatedapplication.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

Inthisstudy,weinvestigatetheproblemofsocialbiasdetectioninonlinetextual content by designing and evaluating a transformer-based system tailored to real- world data. We focused on identifying multiple forms of bias, including gender,caste,religionandraceusingacombinationofbenchmarkdatasetsand curated examples collected from platforms such as twitter and LinkedIn. Throughout our experiments, we observed that the proposed approach is effective in detecting explicit and strongly expressed bias , while also maintaining competitive performance on standard evaluation datasets.

At the same time, our analysis revealed important limitations, particularly whendealingwithsubtle,context-dependent,orunderrepresentedformsof bias such as caste-based discrimination and microaggressions in professionalsettings.

These findings reinforce the need formore balanced, diverse,andcontext-awaredatasets,aswellasevaluationstrategiesthatgo beyond traditional hate speech benchmarks. Overall, our work highlights boththepotentialandthecurrentchallengesofapplyingtransformer-based modelstobiasdetectionandundeerscorestheimportanceofcontinued research towars more inclusive and socially responsible NLP systems.

B. FutureWork

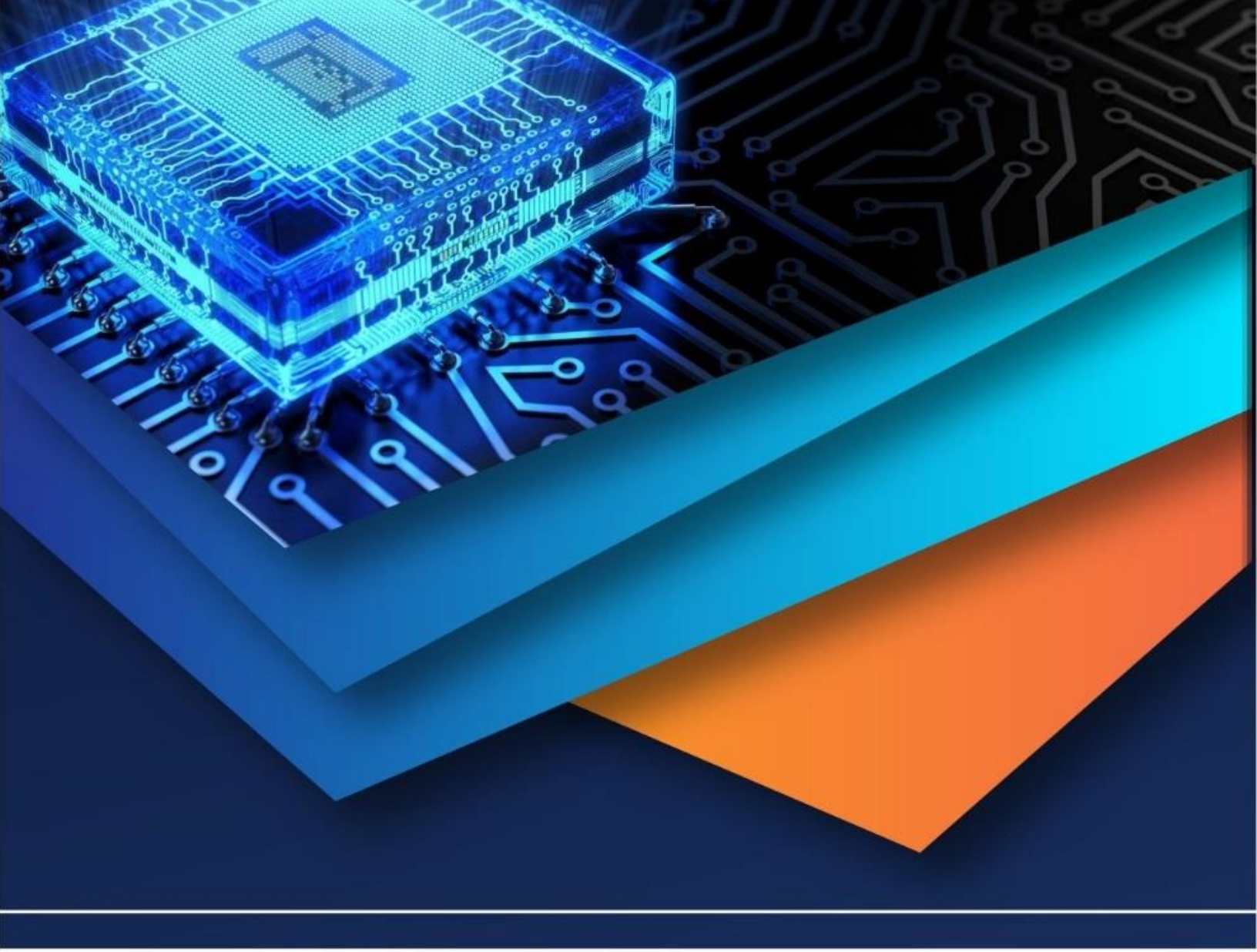
We hve identified five specific areas where this architecture can be expanded.Building on the findings of this study, we plan to extend our work inseveral important directions. First, we aim to apply the proposed bias detection system in practical settings such as social media platforms, where it can assist in real-time content analysis and moderation. We also see potential in integration the model into intelligent text input systems, including virtual keyboards and writing assistants, to provide bias-aware suggestions during text composition. In addition,we plan to explore its use within search engines , where identifying andreformulating biased queries could contribute to more neutral and inclusive informationaccess.

VII. ACKNOWLEDGEMENT

Wewouldliketoexpress oursinceregratitudetooursupervisor, Mr. Jagbeer Singh,forhisguidanceand constructivefeedback throughoutthisproject.His support helped shape the direction and execution of our research. This work representsthecollectiveeffortofourteam-NityamKumarTiwari,Neeraj Patel,HimanshuandNishthaGarg-whocollaborativelydesigned, implemented and evaluated the Ethical AI Analyzer system.
WealsothanktheDepartmentofComputerScienceandEngineering,MIET,forprovidingtheacademicenvironment,infrastructure,andresearch supportnecessarytocarryoutthiswork.Finally,weacknowledgethe publiclyavailabledatasets, openresearchresource, andpriorstudiesin Ethical AI thatinformedandstrengthenedourapproach.
Computer Science and Engineering at MIET for giving us the resources and lab supportweneeded. Finally,weacknowledget hedevelopersbehindtheGemini, Hugging Face and OpenAI APIs which made our experiments possible.

REFERENCES

[1] A.Founta,C.Djouvas,D.Chatzakoa,L.Tiadis,J.Blackburn,G.Stringhini,A.Vakali,M.Sirivianos,andN.Kuortellis,"Largescalecrowdsourcingandcharacterization ofTwitterabusivebehavior,"Proc.Int.AAAIConf.WebandSocialMedia,vol.12,pp.491-500, 2018.
[2] A.Field,S.L.Blodgeett,Z.Waseem,andY.Tsvetkov,"Asurveyofrace,racism,and anti-racism in NLP, " in proc. 59thAnnu. Meeting Assoc. Comput. Linguistics (ACL)pp.1905-1925,Aug.2021.
[3] T. Bolukbasi, K. -W. Chang,J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmeras women is to homemaker? Debiasing wordembeddings, "AdvancesinNeutralInformationProcessingSystems,vol.29,pp.4349-4357,2016.
[4] C.Basta,M.R.Costajussa,andN.Casas,"Evaluatingtheunderlyinggenderbiasincontextualizedwordembeddings,"inProc.1stWorkshoponGender Bias in NLP, pp. 33-39, 2019.
[5] W. Guo and A. Caliskan, "Detecting emergent intersectional biases: Contexalzedwordembeddingscontainadistributionofhuman-likebiases,"Proc. ACM Conf. Fairness, Accountability, and Transparency, pp. 122-133. 2021.
[6] M. jiang and C. Fellbaum, "Interdependencies of gender and race in contextualizedwordembeddings,"inProc.2ndWorkshoponGenderBiasinNLP,pp.17-25,Dec. 2020.
[7] Y. C. Tan and L.E. Celis, "Assesing social and intersectional biases in contextualized word representation, "arXiv preprint arXiv: 1911.01485, 2019.
[8] S.Sharifirad,A.Jacovi,andS.Matwin,"Learningandunderstandingdifferent catagories of sexism using convolutional neural network filters, "in Proc. Widening NLP Workshop, pp. 21-23, 2019.
[9] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, and V. Verma, "Multi-label categorization of accounts of sexism using a neuralframework, "in Proc. EMNLP-IJCNLP, pp. 1642-1652, 2029.
[10] H.Liu,W.Wang,Y.Wang,H.Liu,andJ. Tang,"Mitigatinggenderbiasfor neuraldialoguegenerationwithadversariallearning,"inProc.EMNLP.
[11] H. Liu, W. Wang, Y. Wang, H. Liu, and J. Tang, "Mitigating gender bias for neuraldialoguegenerationwithadversariallearning,"inProc.EMNLP,pp.893-903,Nov.2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)