



AI-Driven Foul Message Flagging

Prof. Aniket Kore^{1*}, Simran Bardhan², Rishita Merchant³, Yogesh Jha⁴

¹Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra aniket.kore@djsce.ac.in

²Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra simranbardhan13@gmail.com

³Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra merchant.rishita2912@gmail.com

⁴Dept. of Computer Engineering DJ Sanghvi College of Engineering Mumbai, Maharashtra jhayogesh30@gmail.com

Citation: Prof. Aniket Kore (2024), AI-Driven Foul Message Flagging, *Educational Administration: Theory and Practice*, 30(5) 14810-14816

Doi: 10.53555/kuey.v30i5.7446

ARTICLE INFO

ABSTRACT

The surge in online communication has posed challenges in maintaining civility, prompting the integration of AI for foul language detection. While AI can effectively identify offensive language, its limitations with sarcasm and slang necessitate a hybrid approach with human oversight. This paper advocates for configurable filters that blend AI's computational power with human expertise to adapt to diverse online communities. Through a review of existing literature and case studies, it proposes guidelines for transparent and accountable AI-powered flagging systems, emphasizing ongoing collaboration between AI and human moderators. This research contributes to fostering a cleaner and more civil online environment by acknowledging the need for both technological advancement and human intervention in mitigating offensive content.

Keywords: Natural Language Processing (NLP), Joint-Learning models, Multilingual BERT, Cross-Lingual hate speech detection, AI-Driven content moderation, Foul language flagging, Linguistic sensitivity, Digital safety, Ethical implications.

I. INTRODUCTION

The first critical aspect of this endeavour lies in hate speech detection. Hate speech, characterized by its discriminatory or derogatory nature towards individuals or groups based on attributes such as race, ethnicity, gender, religion, or sexual orientation, poses a significant threat to online discourse and community cohesion. AI-powered algorithms, trained on vast amounts of text data, can effectively identify patterns associated with hate speech, enabling automated systems to flag and remove such content promptly. However, the effectiveness of hate speech detection algorithms relies heavily on the quality and diversity of the training data, as well as the ability to adapt to emerging forms of discriminatory language.

Another challenge in foul message detection pertains to the nuanced nature of communication, particularly in discerning sarcasm. Sarcasm, often employed as a form of humour or satire, can confound AI algorithms due to its subtlety and context-dependent interpretation. Current AI models struggle to accurately detect sarcastic remarks, leading to potential misclassification of benign content as offensive. Addressing this challenge requires advancements in natural language processing (NLP) techniques and the integration of contextual clues to enhance the algorithm's understanding of sarcasm within different cultural and linguistic contexts.

Finally, the effective flagging of offensive messages necessitates a balanced approach that combines AI-driven automation with human oversight. While AI algorithms can efficiently detect explicit profanity and some forms of hate speech, they are not infallible and often overlook the nuances of language and cultural context. Human moderators play a crucial role in refining and calibrating AI-powered flagging systems, providing context-specific insights and addressing instances where automated detection falls short. By integrating human judgment with AI capabilities, online platforms can develop configurable filters that adapt to the unique needs and preferences of diverse communities, thereby fostering a more civil and inclusive online environment.

II. REVIEW OF LITERATURE

The paper labelled as [1] focuses on developing an online hate classification system for multiple social media platforms. It thoroughly explores dataset composition and labelling methods, employing various machine learning algorithms such as LR, NB, SVM, XGBoost, and FFNN. Feature representation includes Bag of Words, TF-IDF, Word Embeddings, and BERT. A central aspect is the use of a large, labelled dataset for abusive language detection, although the specificity to Twitter and subjective human annotation are acknowledged as limitations. Nevertheless, the paper provides valuable insights into the construction of robust hate classification systems suitable for diverse social media landscapes.

The literature review, referencing paper [2], provides a comprehensive analysis of the challenges in distinguishing offensive language from hate speech online. While it recognizes the paper's valuable contribution in addressing this issue, particularly through the introduction of a novel dataset from Twitter, it also notes potential limitations such as platform specificity and biases introduced by human annotation methods. Despite these concerns, the review acknowledges the significance of the dataset's diverse content and the insights gained from employing machine learning algorithms alongside human annotation. Additionally, the review highlights specific obstacles in differentiating hate speech from profanity, emphasizing the complexities involved in automated identification and the nuanced nature of offensive language. Overall, it underscores the importance of ongoing interdisciplinary research efforts to effectively combat online hate speech and promote a safer digital environment.

The literature review, attributed to paper [3], presents a comprehensive survey on troll detection methods, offering insights into existing approaches and the prevailing challenges in the field. Despite the absence of a specific dataset mentioned, the review provides valuable background information on the complexities associated with troll detection, outlining the various obstacles researchers encounter. By summarising existing methods, the paper offers a consolidated view of the current landscape of troll detection techniques, facilitating a better understanding of the methodologies employed in this domain. Moreover, the review highlights the ongoing challenges faced by researchers, shedding light on areas that require further investigation and innovation. Overall, the paper serves as a valuable resource for researchers and practitioners alike, offering a holistic overview of troll detection methodologies and paving the way for future advancements in the field.

The literature review, citing paper [4], examines hate speech detection models, emphasising their multilingual capabilities and employment of zero-shot learning with LASER embeddings. Despite recognizing challenges like limited data for low-resource languages, paper [4]'s contribution in assembling a multilingual Twitter dataset and its innovative approach of joint learning with knowledge injection are praised. Overall, the review underscores the importance of developing reliable hate speech detection models for diverse languages and highlights the potential of collaborative learning strategies to address this task effectively.

The literature review, attributed to paper [5], offers a comprehensive evaluation of resources and benchmark corpora available for hate speech detection on social media platforms. By examining various annotated corpora, benchmarks, and lexica, the review aids researchers in selecting suitable datasets for their studies. Following Kitchenham's guidelines, the review rigorously searched Google Scholar and Books for relevant English and multilingual content across two distinct periods, ensuring a thorough analysis of available resources in the field.

Paper [6] conducts a targeted analysis of tweets surrounding George Floyd's death in May 2020, with a specific focus on French tweets. Employing annotation on Yandex Toloka, the study develops French hate speech detection models leveraging multilingual BERT, CamemBERT, and transfer learning with HateXplain. By categorizing tweets into hate, offensive, or normal categories, the research underscores the importance of multilingual racial hate speech detection using transfer learning. Demonstrating the efficacy of this approach across languages, the study contributes valuable insights into the detection and mitigation of hate speech in diverse linguistic contexts.

The paper cited as [7] explores predictive features for detecting hate speech on Twitter, with a specific emphasis on distinguishing between hateful symbols and individual expressions. Drawing from a dataset sourced from Twitter, the study investigates characteristics that are indicative of hate speech, offering valuable insights into the complex dynamics of online discourse. By focusing on this differentiation, the research contributes significantly to the understanding of hate speech detection, providing nuanced perspectives essential for researchers and practitioners engaged in computational social science.

In paper [8], the focus lies on exploring user network features for hate speech detection on Twitter and assessing different model architectures for combining embedding types. With a dataset comprising 16,000 tweets labelled as racist, sexist, or neutral, the study compares deep learning architectures (including FastText, CNNs, LSTMs) with baselines such as character n-grams, TF-IDF, and Bag-of-Words vectors. Evaluation metrics such as precision, recall, and F1 score are employed to gauge model performance. The research aims to advance hate speech detection on Twitter by leveraging deep learning techniques to achieve superior accuracy compared to existing methods.

Paper [9] acknowledges the challenges in automatic aggression detection and highlights the significance of further investigation in this area. With a dataset comprising 18,000 tweets and 21,000 Facebook comments, it serves as a valuable resource for researchers. The study collected and labelled Hindi-English social media

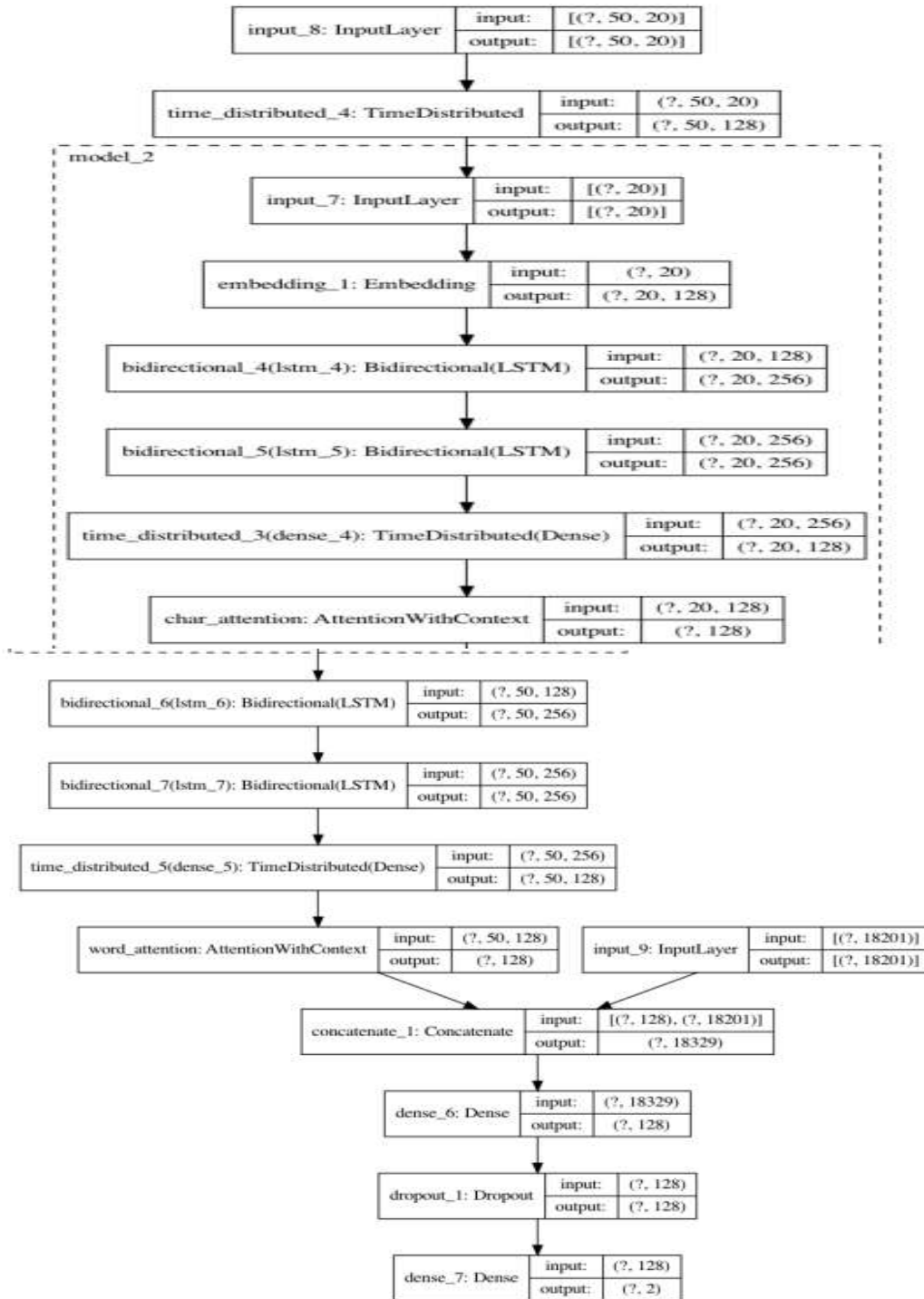
comments, focusing on Indian topics, annotating them for aggression type and effect. An improved annotation scheme was implemented to address initial discrepancies, resulting in higher agreement levels and enhancing the quality of the aggression-annotated corpus of Hindi-English code-mixed data.

In the study cited as [10], novel deep learning techniques are explored to effectively capture textual cues related to hate speech, alongside methods for identifying sarcasm and misinformation within hate speech. Leveraging a public dataset of Hindi-English code-mixed tweets, the research also employs a larger dataset to train domain-specific word embeddings. Three deep learning models—CNN-1D, LSTM, and BiLSTM—are implemented for hate speech detection, with their performance compared to previous statistical approaches like SVM and Random Forest. Through rigorous evaluation using 10-fold cross-validation on a benchmark dataset, the study contributes valuable insights into hate speech detection from code-mixed Hindi-English tweets, offering promising avenues for future research in this domain.

III. METHODOLOGY

A. Architectural Diagram

Fig. 1. Architectural Diagram



B. Proposed Work

AI-driven foul message flagging aspires to cultivate a more civil and inclusive online environment. By shifting through mountains of text, AI can identify a spectrum of offensive language, from hate speech spewing venom to everyday swears that grate on some users. This not only improves the overall user experience by fostering a more respectful online space but also frees up human moderators to tackle complex situations where AI might stumble. Additionally, AI can categorize content based on language, ensuring a safer experience for younger users or those sensitive to offensive language. This system also offers a crucial advantage in the fight against online negativity – the ability to continuously learn and adapt to new forms of offensive language, including slang and evolving hate speech tactics. By combining AI's detection prowess with human oversight, online platforms can strive toward a cleaner, more positive digital space. Our system boasts a robust, multilingual dataset (social media, forums) enriched with data augmentation for cultural understanding. Transfer learning and cross-lingual NLP techniques ensure seamless handling of diverse languages. Advanced NLP and machine learning minimize false positives/negatives, effectively addressing

sarcasm and ambiguity in text. Designed for scalability and real-time monitoring, the system utilizes precision-driven NLP for accurate detection.

C. Implementation

A) Datasets

The dataset comprises Hinglish tweets and comments meticulously annotated for hate speech, derived from a plethora of social media platforms and online communities. Through detailed categorization of hate speech instances, the dataset provides a comprehensive foundation for hate speech detection. To bolster sarcasm detection capabilities, data augmentation techniques were strategically employed, resulting in the generation of additional instances of sarcastic tweets and comments within the Hinglish hate speech dataset. These augmentation methods, including paraphrasing and context modification, were intricately designed to diversify the training data, thus facilitating exposure to a broader spectrum of sarcastic expressions. The augmented dataset functions as a pivotal component of the training pipeline, with the overarching goal of refining the model's sarcasm detection performance. By incorporating this enriched dataset, the model is poised to accurately identify sarcastic content embedded within Hinglish text across an array of online platforms and forums, thereby significantly enhancing its utility and effectiveness in real-world applications.

a) Aggressive Content Dataset

Fig. 2(a). Trained Processed Aggression Dataset

```

well said bano..you have courage to stand against dastgiri of Muslims OAG
Peak of Private Banks ATM's Like HDFC, ICICI etc are out of cash. Only Public sector bank's ATM working NAG
How question is, Pakistan will adhere to this? OAG
Pakistan is comprised of fake muslims who does not know the meaning of unity and imposes their thoughts on others.....all the rascals have gathered there... OAG
Phe r against cow slaughter,so of course it will stop leather manufacturing if it happens. NAG
Wondering why educated Ambassador is struggling to pay through Credit/Debit at a Decent Restaurant! Cest imagine that diplomat of a Developed nation is not having a Card and he needs Cash only for Dinner. CAG
How does inflation react to all the after shoots of this demon...? NAG
Not gone job....this guys creating a problem a our society CAG
This is a false news India media is simply misguiding there nation and creating hatred..Media should be v careful while spreading the news..SHAME..! NAG
No permanent foes, no permanent friends. Interest is permanent ! NAG
Deepak Kumar Sharma Saab...chalo ap ki Ya baat ek baar marn li...Now whatever pr Saab has talked about in 2016...Kya kya Kis us main...He is just a campaigner of BJP nothing else,ek cheer bta do.... but he is implementing all the bills which he opposed before 2016.. Example- FDI,GST and list is long... NAG
Communist parties killed lacks of opponents in WB in 35 years ruling????? OAG
Why you guys counter the medi govt decisions, its fact there is black money cleanup stand taken ... so many discussions on news channel, individuals meetings why no such efforts made by media and make people scared and provoke" NAG
Wt is 3 time ban terrorist organization OAG
No
Same acting ll be there ...
but we ll watch so option !!!
👍👍👍 NAG
Happy Diwali!!!let's wish the next one year health,wealth n growth to our Indian economy. NAG
lol... He said he is gonna employ large number of people in cyber security sector.. are you gonna employ illiterates in that profile???? NAG
So funny stupid,,,,,, OAG
absolutely! the deeper you give the shallower customer you have. CAG
From Saheb , aeti national leftist comies , media and Muslim pandering Hindu sick-o-lars have milwed ""secularism"" to all it was worth to render it into a mere failed slogan . Sics-O-larism has been nothing more than continuation of British policy of ""Divide and rule"" . That RSS is opposed to ""the idea of secularism"" , absolutely because it has been used as a dagger into the heart of hindus , hinduism and Hindustan . CAG
You nifty above 20 day moving average, what next 30 day moving average BISS NAG
Good to see when their in so terror in pak and afgan and india as well so in can fur development step NAG
    
```

Fig. 2(b). Validation Processed Aggression Dataset

```

The quality of re make raw rates or TRIM it is something to be bought from fish market CAG
"ative
How is ur mother???
How is ur wife???
How is ur sister!!!

Hope everyone is fine...👍👍👍 NAG
Also see ...hu ur RSS activist caught in Buxtha .... throwing beef in d holy temples...https://www.google.co.in/amp/www.india.com/news/india/buxtha-clad-rss-activist-caught-throwing-beef-at-temples-pictures-go-viral-ct-facebook-69316/amp/ NAG
The death of 2 jians in LOC CROSS FIRE!!!
OUR H'BLE HOME MINISTER MUST SING;

KANG: KADI NENDA.....
SARKARA: GO:
(UTAR) RAJ HATH SIMHA & APCEY SAHA. NAG
Nest ho ya Nandohar sirgh saala yeh dog kash re karta hai sirf note ka Lia De ha saara dikhatata hai OAG
The discrimination in policies and non protectionism is perhaps the Fundamentals of SECULARISM. CAG
I m reservation on the basis of caste Religion and communities to understand the essence of SECULARISM. CAG
As Mr Ashwini gujaral said once nifty will touch 2200 , it is most likely nifty will test 2200, 2100, 2070, & 2050 NAG
After seeing your comment guys... I think conu slogan is right cause you guys can only troll all religions. CAG
what has so far Mr.Nehru done for this country. Ask him to shut down his bloody pistols for good or I ll given the chance will trap at his mouth hole. CAG
SENIA SHINDY..... A COMMENT.....WRONGLY INTERPRETS HER AS A NARRATOR.....TA YGAS.....IN A SERIOUS BUSINESS CHANNEL SHE ACTS AS IF SHE IS IN A ENTERTAINMENT CHANNEL..... PLEASE DONT SHOW YOUR UCHA UCHA DACT.....LEARN IT FROM ESTA BAPPA AND WITH TECHNOLOG..... CAG
Now all AAPPrads will abuse ANNA HAZARE NAG
I salute ..@ keel Patel..i r just smiling. back & every comment of urs is true & correct...India a world need people like u....love u my brother..God bless u. & pls don't jstg here. Keep ur comments on every required post... NAG
No Sonia! Harbets & reliance sounding like rocket, what's your base re nifty breaking all time high NAG
These kamardi youths are radicalized jikad terrorists. There are a few exceptions, but most of them are religious fanatics so keeping them at home is equal to keep a time bomb at home. OAG
If you want to construct a smooth road you have to remove garbage and it's difficult. If you want go on highway you have to come out from street. Some problems are necessary to enter in next class. 🙄👍👍👍 IF CAG
    
```

b) Hate Speech Dataset for Hinglish

Fig. 3. Hate Speech Dataset for Hinglish

index	text	category	text length
0	I am Muslim. Aur main bhi sahi se Pally Pakistan hu. agr tu sahi khatr jisse badshah ki is zameen ki behumat kare aur abhi kar a saro phero Durr chahiye. Proud to be a Muslim and Pakistani	no	186
1	Doctor sab sahi me ke PhD in fake politics? kaha Etha parha bhre ho fr kya jo sab baad kaha ho. Tum bas booring khalo. aur majhe jo pic better conthi qdQDare	no	166
2	Power Death me Pate (DCC me aare Pate) koi gupat Ki oter kar may be, ye manavallyon baansanahi kabhi saplo anarichan kahi deyo je te jo (DCC Ki Mia) hai use hi nahat kaha hote je khooor aur chandil ka hok kama usake bhramhanaradil ki ke saps nahat hote	no	257
3	Sarkar banne ka bad Hindu bhime ek teli bhala Jo kyo ke dvara hai gaya hu kyo ke gar gubar mandir mangit aur nahat kaha kar ret chahiye	yes	148
4	Hate fr jst not ate hu pr hi kaha me ho --	yes	45
5	Et off ek paar ake ki embolan ki K? ke rape aur booring between hoi lega after rape honge justified due to chomach. bhery naj toh Rajon ki ho gaye aban ante ki gusan	no	174
6	Alas lego na sahte nahat kaha hu Jo raste ko nahi kar sahe jst ke chaudi hote ki but nah me ve paku hote fr	no	166
7	Ok jyada sentimental mat ho jar peony ke baad. Das death ki booring off Hindu nahat pe hai ho woh kahi nahat vedega. Aap FOK de do ya phir Kallantar ake. jwaki off gahen se ellega	yes	191
8	I am very sorry to say saad ki staha ke bane hote ke ki main kya hai woh hi sampta hain hain hoi kar nahat ki kin	no	121
9	Muslim Bahut achha hai. Hii CSS jary ye jary ki malim me jalkucha from ki. Aal sahe "MILIM" hantel hai. Jo kahi Parton Pato Tempa. GaAlata (BhaadMastaki. Vardelitaran. Nasatragan. Pao Pao. Gaddar Baidhe jyah. SBOyQSBahche ki sahar hai kaha Hany "LUMINAL" j jaha ake hai	no	286
10	I hate anyone who Call them to kin chakana add me aur tum agar hai chhalenge toh kama shro. bhare se chakre. https://twitter.com/akshayvishnu/2518860523266	yes	176
11	Ha san "emoji" hote hii emoji? Mujhe nahat ki hi pasd aati hai	no	60
12	Hate hene... the much booring n kar kama ka kama ohi hai usake... sate ki small kama jst hai usake... hoi aur kach nri ad hai... jstot booring hai wo kahi	yes	178
13	Ok agr hi kar hai ho kama ma?	no	36
14	Mehant to school me copy kaha bhre me ki lego hai fr lego kyo? Atanki ko ki lego hai fr nahat kyo? Pate kaha ghar ki aare ki anam bech doya kya pake aur mehanat ka kya aap?	no	189
15	Ladke wo hai jo kar jom kay bad hii ghar nahat hote manay mat nahat dera aur kama nahat kahi. hi number kaha kay aur phara nahat lego	no	142
16	aur ye kama ka comabk show hai. hai ki nahat dekha chahiye? kahi fr ki nahat dahi gaye... mat bhavni ki q hote kary? meza uski aching nahat puchi legi je meza lego vo make kaha. wo meza apden hai	no	253
17	Se j ab ye possible nahat hai nahat nahat hai dera on ab ek ki kama kar nahat ke kahi ko mila kar hi ye khamar ke kaha hai jama nahat ki me gaye on nahat ki hii ab fr se ek kary hai dera mu ki (Chomach) anahat bhavni ek hii ake hii. bhare comabk ap rasta/2N42458185091252	no	268
18	Code phanta hai toh phanto de. Mujhe project se nahat ho gay hai. Mujhe development team se hantre de	yes	162
19	Anti hater jani chahiye... I hate and khate	yes	44
20	Acha bhi dya ham n show hi nri dekha the. Kahi bhi kach on ant nahat hoi ki meza to hote	no	96
21	Karachi me apneve pengi. parham ki qal kelay anti nahat gese bhoyi hain kaha me nahat kay kahi nahat pake ho	yes	126
22	Kacha ke Hindu bhramhanaradil au nahat hai kar	no	44
23	aj anhi me je kar vake pichade me nahat fr me phote hok ki nahat jst ok	no	79
24	Ar day me ki kama on hii hii ke mehi me hantel dahi chanta ki sa kar kama anahat ki hii	no	52

c) English and Spanish Hate Speech Dataset

Fig. 4(a). English Hate Speech Dataset

id	text	HS	TR	AG
1	RT @vaantshil: shut the fuck up and come suck my dick	1	1	1
2	@ArianasBotch OK if you fucking said leave black me. But d m me first i'm gonna kick your ass. Shut the fuck up you https://t.co/SBkMfMEt9	1	1	1
3	@Cyy_Sw Wow me cock got hard. Went to pull your panties down, push you against these poor and let you feel my cock deep in your wet pussy aT	1	1	1
4	It will be bitch (chloe) when your not home OK? idgal if you ok?e kidding bitch lol?i report your ass... https://t.co/SBkMfMEt9	1	1	1
5 I get to rape beautiful women and that's why I like working on this project.	1	0	1
6	Onke Mercedes to Paul Schabas, "Chae's a scoww. I attack Indigenous women all the time. No one will notice if you rape her! Rape her!"	1	0	1
7	lol chop her head off and rape the bitch https://t.co/2F8Coun5D8	1	1	1
8	@Hao0R0: I love your ass bitch https://t.co/vy0im1TS7	1	1	0
9	@hasuka89 go rape your whore of a mother!	1	1	1
10	If she lets you put your penis in her vagina with no strings attached, have the common decency to treat her like a queen on her birthday	1	0	0
11	WOW can't believe all these women riding the subway today? Shouldn't these bitches be making sandwiches LOL @stefemales.	1	0	0

Fig. 4(b). Spanish Hate Speech Dataset

id	text	HS	TR	AG
101	@ian_delaCalle @IrantzuVarela @pikaramagazine Oye por quÃ© no molestas a tu puta madre?	0	0	0
102	NINGUNA MUJER ES PUTA! âââ https://t.co/cVOCQMSRlg	0	0	0
103	Editar, ademÃ¡s de complicado, es lo que hace de algo que merezca la pena o no, porque puedes haber grabado la puta hostia de plomo, que si no los montas bien, con un sentido y los maquilas atractivamente, se va a la mierda.	0	0	0
104	Bien joder una puta a legrÃ¡ all Te lo mereces pequÃ©ta #OTGala9	0	0	0
105	Todo polÃ¡tica... no se levanta la sesiÃ³n y estos hijos de puta mandan a los peones que estÃ¡n en la plaza y provocar violencia... 15 policÃ¡as heridos pero no pasa nada... si esos que estÃ¡n ahÃ¡ son el pueblo... DIOS MIO el pueblo es mi vieja que se rompe el lomo laburando no ellos.	0	0	0
106	Casuso callate de una puta vez https://t.co/iWmjWlUFMM	0	0	0
107	@BrdMM @NayBetania TÃ¡ eres la culpable, perra >>>	1	1	1

d) Sarcasm Content Dataset

Fig. 5. Processed Sarcasm Hinglish Dataset

Text	Category
Triple Talaq par Burbak Kuchh nahi bolega	NO
Batao ye uss site pr se akki sir ke verdict nikaal laaye jaha he aaj ki ek bi movie hit nai	YES
Hindu baheno par julam bardas nahi hoga @TripleTalaq Hindu daram par lago hoga hamari Hindu baheno ki soraksa ke liye	NO
Naa bhai.. aisa nhi hai.. mere handle kame se bhi kuchh hona nhi hai.. politics se mera door door tak ka naata nhi hai	NO
#RememberingRajiv aaj agar musalman auralen triple talaq ki wajah se itni jaleel ho rahi hai to iss ka sabse bada reason Rajiv gandhi hai.	NO
are cricket se sanyas le liya kya viru aur social service suru kardiya.kheli hi bhul gaye.2 innings 0 n 0	NO
Sohail bhai hame bhi treen kardo muje bhi shock he cricket	NO
bhai triple talaq se aap kya samjhle hai samjhaye aap zara .agar triple talaq pta hota apko loh aisa nhi kehte..	NO
Bachcho ki death par politics ke bajay unke liye prayer aur unki family ko console kiya jana jyada zaruri he.. thodi humanity bhi zaruri he	NO
#@spn_cricket ijiye S Thakur ne phir pahni no.10 jarsi, inka kharab din chalu ho gaya..@BCCI#Sisvind	NO

B) Models

1. **Data Collection and Augmentation:** The foundation lies in data. We collect multilingual data from various online communities (social media, forums) encompassing both positive, negative, and sarcastic text. To account for cultural nuances and expand dataset size, we employ data augmentation techniques. This can involve techniques like synonym replacement, back-translation, and random shuffling to create variations of existing data points, fostering a more robust and generalizable model.

2. **Text Classification with Ensemble Learning:** The core of the system lies in utilizing multiple models for text classification.

(i) Aggressive Content Dataset

- **Categorical Classification:** The problem deals with classifying text into three categories: aggressive, non-aggressive, and neutral. This suggests a multi-class classification problem.
- **Softmax Activation:** The final layer of the neural network uses a "softmax" activation function. This function is commonly used in multi-class classification problems to convert the network's output into probabilities for each class.
- **Categorical Cross-Entropy Loss:** The code uses the "categorical_crossentropy" loss function. This loss function is specifically designed for multi-class classification problems.

(ii) Hate Speech Detection

- **Hierarchical Attention Network (HAN):** HAN is a neural network architecture designed to handle hierarchical structures in data, such as text. In hate speech detection, HAN can effectively capture the hierarchical relationships between words, sentences, and documents, enabling more accurate classification.
- **Transformer (e.g., BERT):** Transformer models, such as BERT (Bidirectional Encoder Representations from Transformers), are state-of-the-art architectures for natural language processing tasks. In hate speech detection, BERT can leverage its bidirectional context understanding to capture subtle linguistic cues indicative of hate speech.
- **Word-Level Long Short-Term Memory (WLSTM):** A regularization term is part of the Lasso Regression linear regression model. It does both feature selection and regularization by adding the absolute values of the coefficients to the loss function. The less significant features are shrunk towards zero, which promotes sparse solutions. Lasso Regression can be helpful for handling multicollinearity and feature selection. When working with high-dimensional datasets, it is efficient.

(iii) Sarcasm Detection

- **ZSC (Zero Shot Classification):** Zero-shot classification is a machine learning technique where a model is trained to recognize classes it has never seen during training. In sarcasm detection, zero-shot classification can be employed to categorize text instances as sarcastic or non-sarcastic without the need for labelled sarcastic data. By leveraging pre-trained language models, such as GPT (Generative Pre-trained Transformer), zero-shot classification can effectively identify sarcastic expressions based on their contextual understanding.

IV. CONCLUSION

The novel ensemble method proposed for hate speech and sarcasm detection in textual data represents a sophisticated approach that amalgamates several established models, including BERT, HAN, and WLSTM, with a zero-shot model to enhance efficiency. By leveraging pre-trained models, the method ensures heightened accuracy in detecting both hate speech and sarcasm, thereby contributing to a more nuanced understanding of text data. Additionally, the incorporation of data augmentation techniques serves to account for diverse cultural nuances, enriching the model's adaptability across different linguistic contexts. Furthermore, AI-powered flagging mechanisms, such as rule-based filtering and sentiment analysis, are employed to categorize identified text types, enabling the system to take appropriate actions. For instance, hateful content identified by the ensemble method is promptly flagged for removal, while instances of flagged sarcasm are contextualized to prevent misinterpretations and foster accurate comprehension.

V. FUTURE SCOPE

In terms of future scope, expanding the project to detect languages in regional languages presents a significant opportunity to enhance its applicability and inclusivity. By incorporating language detection capabilities for regional languages, the ensemble method can effectively cater to a broader user base, facilitating a more comprehensive analysis of text data across diverse linguistic landscapes.

Additionally, developing a user interface (UI) that supports multilingual languages would further augment the accessibility and usability of the system. A well-designed UI capable of seamlessly accommodating multiple languages would enhance user experience and ensure that individuals from different linguistic backgrounds can effectively interact with the system.

Furthermore, integrating additional features such as aggregation datasets and the analysis of various emotions would enrich the functionality of the ensemble method. Aggregation datasets could provide valuable insights by consolidating data from multiple sources, thereby improving the robustness of the model. Similarly, analysing a wide range of emotions beyond hate speech and sarcasm could offer a more nuanced understanding of text data, enabling the system to detect and respond to diverse emotional expressions effectively.

Overall, these future scopes hold immense potential for advancing the capabilities of the project, ultimately contributing to its effectiveness in promoting a positive and respectful online environment. By embracing language diversity, refining the user interface, and incorporating additional features, the ensemble method can evolve into a comprehensive solution tailored to meet the evolving needs of online content moderation and sentiment analysis.

REFERENCES

- [1] Salminen, J., Hopf, M., Chowdhury, S.A. et al. Developing an online hate classifier for multiple social media platforms. *Hum. Cent. Comput. Inf. Sci.* 10, 1 (2020). <https://doi.org/10.1186/s13673-019-0205-6>
- [2] Shervin Malmasi & Marcos Zampieri (2018) Challenges in discriminating profanity from hate speech, *Journal of Experimental & Theoretical Artificial Intelligence*, 30:2, 187-202, DOI: 10.1080/0952813X.2017.1409284
- [3] Tomaiuolo, Michele & Lombardo, Gianfranco & Mordonini, Monica & Cagnoni, Stefano & Poggi, (2020). A Survey on Troll Detection. *Future Internet*. 12. 31. 10.3390/fi12020031.
- [4] Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection, *Information Processing & Management*, Volume 58, Issue 4, 2021, 102544, ISSN 0306-4573,
- [5] Poletto, F., Basile, V., Sanguinetti, M. et al. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang Resources & Evaluation* 55, 477–523 (2021). <https://doi.org/10.1007/s10579-020-09502-8>
- [6] Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023. Multilingual Racial Hate Speech Detection Using Transfer Learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 41–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- [7] Talat, Zeerak and Dirk Hovy. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." *North American Chapter of the Association for Computational Linguistics* (2016).
- [8] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Investigating deep learning architectures for hate speech detection on Twitter.
- [9] Kumar, R., Reganti, A. N., Bhatia, A., & Maheshwari, T. (2018). Aggression-annotated Corpus of Hindi-English Code-mixed Data.
- [10] Kamble, S., & Joshi, A. (2018). Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models.