Cyberbullying detection using BERT for Telugu Language

1st Sri Lakshmi Talasila
Computer Science and Engineering
Prasad V. Potluri Siddhartha Institute of Technology
Vijayawada, India
tslakshmi@gmail.com

2nd Dharani Priya Kothuri
Computer Science and Engineering
Prasad V. Potluri Siddhartha Institute of Technology
Vijayawada, India
kothuridharanipriya@gmail.com

3rd Savithri Jahnavi Manchiraju
Computer Science and Engineering
Prasad V. Potluri Siddhartha Institute of Technology
Vijayawada, India
jahnavimanchiraju10@gmail.com

4th Mutyala Sai Sasank Mallavalli
Computer Science and Engineering
Prasad V. Potluri Siddhartha Institute of Technology
Vijayawada, India
sasankmallavalli229@gmail.com

5th Lourdu Gnana Harshith Dande
Computer Science and Engineering
Prasad V. Potluri Siddhartha Institute of Technology
Vijayawada, India
harshithcs28@gmail.com

Abstract- The rapid proliferation of online communication has introduced cyberbullying as a significant concern affecting individuals' well-being. Existing research employs various techniques like Tf-Idf, XLM-RoBERTa, and machine learning algorithms such as logistic regression, SVM, Random Forest, and Naive Bayes to detect cyberbullying across mixed and bilingual languages. However, these approaches often struggle with accuracy and fail to effectively discern cyberbullying instances due to language nuances and context misinterpretation. Key challenges faced by previous systems include limited linguistic coverage, contextual understanding, and nuanced interpretation of cyberbullying. To address these, our study introduces a novel approach utilizing the Indic-bert model tailored for Telugu. BERT (Bidirectional Encoder Representations from Transformers) inherently addresses these challenges by leveraging bidirectional context understanding, allowing it to capture subtle linguistic nuances and contextual cues, thereby improving accuracy and contextual understanding. By focusing on contextual nuances, our model aims to improve cyberbullying detection's precision and effectiveness for Telugu content. We present a Telugu dataset comprising 27,000 sentences and achieve an accuracy rate of 90%, highlighting the efficacy of our approach in overcoming these challenges and contributing to online safety.

Keywords*- *Cyberbullying, Telugu,* Bidirectional Encoder Representations from Transformers (BERT), *Bullying Preprocessing, Harassment, Language, Social Media

I. Introduction

The advent of the internet and rapid technological advancements has undeniably transformed the way we live, communicate, and interact with the world. This digital revolution has connected people across the globe, facilitated instant communication, and provided unprecedented access to information. As our lives have become increasingly intertwined with the virtual realm, the positive impacts of technology are evident in education, business, and social connectivity. However, alongside these transformative changes, a dark underbelly has emerged – Cyberbullying. The digital landscape, once heralded as a beacon of connectivity and information sharing, has also become a breeding ground for harassment, intimidation, and abuse. The ease of access to online platforms, social media, and negative behavior. Cyberbullying, the malicious use of electronic communication to target and harm individuals, has become a pressing concern in this brave new digital world. This transformation has not only altered the dynamics of human interaction but has also brought about new challenges in maintaining a safe and inclusive online environment. As we grapple with the implications of our connected world, it becomes imperative to explore the multi-faced nature of cyberbullying, its impact on individuals, and the necessity for proactive measures to detect and prevent these digital aggressions. This exploration will shed light on the evolving landscape of technology and its darker repercussions, urging us to confront the urgent need for effective cyberbully detection methods in order to preserve the positive potential of the digital era. Cyberbullying is the term used to describe bullying that takes place online. Mobile gadgets, gaming platforms, social media, and messaging apps can all be put to use for it. It's a pattern of behavior designed to spook, infuriate, or humiliate the target audience.

II. Purpose

The purpose of cyberbullying detection is to identify and prevent instances of cyberbullying in order to protect and support the victims and address the behavior of the perpetrators. It involves the use of technology and various strategies to monitor online activities and communication in order to identify potential cases of cyberbullying. By detecting cyberbullying early on, appropriate interventions and support can be provided to the victims, and the perpetrators can be held accountable and educated on the consequences of their actions. It also helps to create a safer and more positive online environment for individuals to interact and communicate. Overall, the purpose of cyberbullying detection is to promote and maintain a culture of respect and kindness in the digital world. The current implementation encompasses Bengali, Urdu, Tamil, and English. However, it's important to note that support for Telugu is not currently available within the system.

To achieve safety in our cyberbullying detection initiative, we're adopting a multi-pronged approach. First, we'll enhance our machine learning algorithms to improve the accuracy and efficiency of cyberbullying detection across languages, ensuring that potential cases are identified and addressed promptly. Second, we'll implement strict data privacy and security measures to safeguard user information, using encryption and secure storage solutions. Third, we'll integrate user feedback mechanisms to continually refine and improve our platform based on real-world experiences and needs. Lastly, we'll promote a culture of digital citizenship through awareness campaigns and educational content, encouraging users to be responsible and respectful online. By combining these strategies, we aim to create a safer, more supportive online environment for all users.

The initiative to develop cyberbullying detection specifically tailored for Telugu represents a crucial step towards fostering a more inclusive and culturally sensitive online environment. By acknowledging the need for anti-cyberbullying measures in languages beyond the commonly supported ones, this project underscores a commitment to global online safety.

III. LITERATURE SURVEY

Natural Language Processing Journal 3 (2023)- This paper's main contribution is the development of a dataset of 12,795 social media texts in the low-resource Tamil language that have been identified as fine-grained abusive speech. Together with the machine learning models, they have experimented with several feature extraction techniques and discovered that TF-IDF and BoW perform better than alternative feature extractors. Moreover, we discovered that the transformer models were especially helpful for BACD and that they performed better for code-mixed text than any other model. (RQ2)

Merits: This paper presents a comprehensive dataset development for the low-resource Tamil language, effectively using traditional feature extraction methods like TF-IDF and BoW. Additionally, the study effectively utilized transformer models, particularly for code-mixed text. ***Demerits:*** The research is limited in scope to the low-resource Tamil language and lacks exploration of advanced machine learning models beyond traditional feature extraction methods.

Toward Detection of Arabic Cyberbullying on Online Social Networks using Arabic BERT Models (2023)- A actual Arabic dataset that has been manually annotated to enhance the data quality is gathered from YouTube and Twitter and used in this paper. For the purpose of ensuring consistency, each experiment underwent three rounds of trials. Numerous assessment criteria, including AUC and the macro F1 score, were employed to assess the classifiers' performance. 84.58% and 85.94%, respectively, are the F1 score and AUC achieved by the best model.

Merits*:* This paper stands out for its high-quality manually annotated Arabic dataset and its consistent experimentation methodology, undergoing three rounds of trials. The study also employs rigorous evaluation criteria such as AUC and macro F1 score.

Demerits*:* The research is confined to the Arabic language, and there is limited comparison with other languages or machine learning models.

Measurement: Sensors 24 (2022)- Information files are produced using the ASDTD class and Social Media Online Natural Language Processing (SMONLP). The ASDTD F-scores for the internal information files are improved to 0.797 and 0.854, respectively, by integrating the command messages from multiple compatible files.

Merits*:* The study showcases improved ASDTD F-scores by integrating command messages and effectively uses ASDTD class and SMONLP for information file production. ***Demerits****:* The research focuses narrowly on specific datasets and methods, limiting its exploration of broader machine learning models.

Rapid Cyber-bullying detection method using Compact BERT Models (2021) – They tuned a variety of tiny BERT models using hate speech data. To address the class imbalance in the data, a Focal Loss function has been implemented. On the hate-speech dataset, they were able to obtain cutting-edge results with 0.91 precision, 0.92 recall, and 0.91 F1-score by employing this method.

Merits*:* This paper excels in tuning BERT models effectively and addressing class imbalance with the Focal Loss function. The study achieved high precision, recall, and F1-score, showcasing its effectiveness. ***Demerits****:* The research is limited to hate speech data and has a narrow focus on specific aspects of cyberbullying, potentially limiting its applicability to broader cyberbullying contexts.

Cyber-Bullying Detection in Social Media Platform using Machine Learning (2021- They studied cyberbullying, its manifestations, techniques, and outcomes, as well as the most current studies on its identification and prevention. They also investigated different types of cybercrime. A total of 35,000+ tweets from Twitter were gathered for the experiment, and the data was cleaned and organized to feed multiple intelligent machine learning algorithms. Five key ML algorithms were then applied to the tweets to classify and predict them into two primary groups: "offensive" and "non-offensive." Last but not least, an analysis of those machine learning algorithms has been done using a number of performance indicators.

Merits*:* This research benefits from a large dataset collected from Twitter and applies multiple machine learning algorithms, leading to classification into "offensive" and "non-offensive" groups.

***Demerits**:* The study is restricted to Twitter data and lacks in-depth linguistic analysis, as well as exploration of advanced machine learning methods beyond classification.

Detect Chinese Cyberbullying by Analyzing User Behaviors and Language Patterns 2019- A Long Short-Term Memory Neural Network-Deterministic Finite Automaton (LND) model is constructed in this research that takes into account not only the language content but also the user's attributes and previous speech on social networks. They employ the data of Douban's reviews by evaluating speech patterns with polarized emotions in the absence of identified content. Next, Chinese cyberbullies' Weibo activities are examined using the newly taught model. Due to the user's behavior attributes and language emotional polarity ratings, the accuracy of cyberbullying detection rises from 89% (using the sensitive lexicon filtering approach) to 95%.

Merits: The research introduces a comprehensive LND model that considers user attributes and emotional polarity, resulting in a significant increase in detection accuracy from 89% to 95%.

***Demerits**:* The study is limited to the Chinese language, and the complexity in model construction may hinder its generalizability to other languages or contexts

IV. PROPOSED MODEL

Our innovative approach harnesses the power of the BERT (Bidirectional Encoder Representations from Transformers) Transformer, specifically tailored for Indian languages known as Indic-BERT. Indic-BERT is optimized for languages like Telugu, making it a robust choice for our cyberbullying detection system. It not only understands the complexities of the Telugu language but also captures its nuances, ensuring a more accurate interpretation of text. In addition to leveraging Indic-BERT's capabilities, our model integrates a multi-layered neural network architecture, allowing for deeper and more intricate analysis of text. This enables the model to capture subtle linguistic cues and patterns indicative of cyberbullying, thereby increasing the detection accuracy. Furthermore, to address the challenge of limited labeled data in Telugu, our model employs transfer learning techniques. By pre-training on a large corpus of diverse data and fine-tuning on a smaller, labeled dataset specific to cyberbullying in Telugu, we optimize the model's performance without requiring extensive labeled data.By combining Indic-BERT's contextual language representations with advanced neural network architecture and transfer learning techniques, we aim to significantly improve the accuracy and precision of cyberbullying detection in Telugu, ensuring a safer online environment for Telugu speakers.

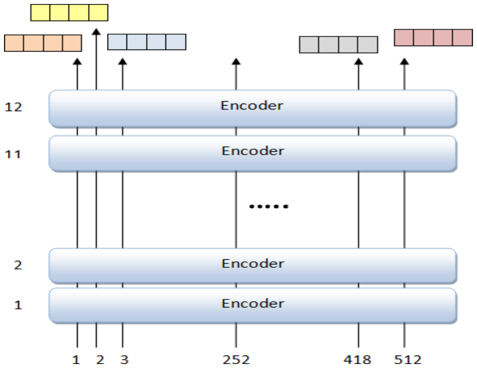


Figure 1: BERT Architecture

V. METHODOLOGY

In our methodology, we introduced a comprehensive model for Telugu cyberbullying classification, incorporating several key steps to ensure the accuracy of the classification process. The journey begins with meticulous data collection, as outlined earlier, wherein we curated a diverse dataset comprising approximately 18,000 sentences covering a range of cyberbullying instances in Telugu. Following this, a rigorous data preprocessing phase was implemented, addressing issues such as missing values and duplicates to refine the dataset.

A crucial aspect of our methodology involves feature extraction, where we harnessed the potential of cutting-edge language models. Notably, we employed the IndicBERT model, tailored to the nuances of Telugu language, resulting in optimal performance for our specific context. The choice of IndicBERT was driven by its ability to capture intricate linguistic patterns and contextual information, contributing significantly to the model's proficiency in understanding Telugu cyber communication.

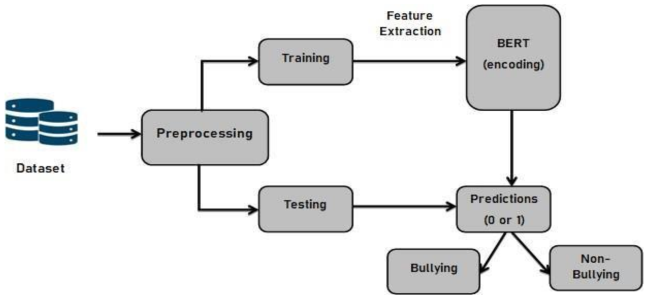


Figure 2: Process Workflow

Furthermore, our proposed model embraces various innovative proposals aimed at enhancing the classification accuracy. These proposals encompass fine-tuning strategies, attention mechanisms, and leveraging domain-specific embeddings to better capture the intricacies of cyberbullying in the Telugu language. By adopting a thoughtful and multifaceted approach in our methodology, we aim to contribute a robust solution to the challenging task of cyberbullying detection in the Telugu linguistic landscape. We provide a brief description of each part's workflow below

i. DATA COLLECTION

Figure 3: Word Cloud

Creating a comprehensive dataset for Telugu cyberbullying detection posed a significant challenge due to the absence of relevant papers or research in the language. Firstly, as research in this cyberbullying aspect in the Telugu language is rare, we couldn't find a source for our dataset. Undeterred by the lack of existing resources, our team undertook the task of crafting a diverse dataset independently. We started collecting data from the comments, tweets, and blogs of various social media platforms. With a substantial collection of approximately 27,000 sentences, our dataset encompasses a broad spectrum of cyberbullying instances, ranging from political and threatening to sexual texts. The dataset is structured with two key columns: the 'Text ' column, containing authentic cyber communication in Telugu, and the 'Label' column, employing a binary classification of 0 for cyberbullying and 1 for non-cyberbullying instances. Throughout the data collection process, careful attention was given to ethical considerations, particularly regarding sensitive content. The dataset not only addresses the immediate need for Telugu cyberbullying research but also holds potential for future investigations in the field. Our commitment to transparency is evident in the documentation of data cleaning, preprocessing steps, and the formulation of training, validation, and testing sets. This lays a robust foundation for the development of accurate machine learning models tailored to Telugu cyberbullying detection..

S.No	Context of Sentence	No. of Text Sentences
1	Bullying	13608
2	Non-Bullying	13812

Table 1: Dataset

ii. DATA PREPROCESSING

In the preprocessing phase of our research, several key measures were implemented to refine the raw textual data, ensuring its quality and suitability for subsequent analysis. Preprocessing has been done for the following:

- **URLs**: We eliminated URLs using regular expressions, effectively removing irrelevant web links to minimize noise in the dataset.
- **User Mentions**: User mentions were generalized by substituting them with the generic tag "@user," fostering anonymity and consistency across the dataset.
- **Numeric Values**: We systematically purged numeric values from the text to reduce potential interference with semantic analysis and streamline the dataset's dimensionality.
- **Emojis**: Emojis underwent conversion into their corresponding text representations, aiding in the integration of non-textual elements for more comprehensive processing.
- **Special Characters**: Characters such as colons and asterisks were replaced with spaces to contribute to the overall normalization of the text.

This step standardized the text and mitigated potential issues during subsequent processing stages. Through these meticulous preprocessing steps, a cleaner and more standardized dataset was achieved, laying a solid foundation for effective natural language processing tasks. Notably, the incorporation of these measures promoted a more efficient analysis process, facilitating the extraction of meaningful insights and patterns from the textual data. Furthermore, the decision not to specifically remove stop words was informed by the utilization of the BERT model, which inherently handles stop words during its tokenization process. Leveraging BERT's advanced language understanding capabilities optimized the preprocessing pipeline for tasks such as sentiment analysis, classification, or language generation..

iii. FEATURE EXTRACTION

Feature extraction is crucial in NLP as it transforms raw text into a format that machine learning models can understand. We chose BERT for feature extraction over traditional methods like TF-IDF or GloVe because of its ability to capture semantic meaning and contextual information, which often leads to more meaningful representations. Unlike TF-IDF and GloVe, which focus on word frequency or co-occurrence statistics, BERT understands the context and semantics of words, making it more suitable for capturing the nuances of language. In our quest to optimize natural language processing (NLP) within our project, the meticulous collection of Bert-related data has emerged as a pivotal undertaking. This strategic decision emanates from a comprehensive evaluation of various Bert models, including mBERT, RoBERTa, and DistilBERT. Each of these models possesses distinct characteristics and functionalities that we carefully scrutinized before arriving at a decision.

mBERT, or Multilingual Bert, is designed to handle multiple languages. While it demonstrates competence in accommodating a wide linguistic spectrum, its performance can be compromised when dealing with languages with specific nuances, such as Telugu. Telugu, being a Dravidian language with unique linguistic intricacies, demands a more specialized model for optimal results.

RoBERTa, an optimized version of BERT, utilizes dynamic masking during pre-training to enhance performance. Despite its advancements, RoBERTa may not fully capture the nuances of Telugu due to its general-purpose nature. The model's pre-training on a vast dataset might not adequately address the subtleties inherent in Telugu, making it less suitable for our project's linguistic requirements.

DistilBERT, a distilled version of BERT, focuses on retaining essential aspects while reducing computational complexity. Although it excels in efficiency, it might not capture the specific linguistic nuances crucial for our Telugu language tasks. The distilled nature of the model may result in the loss of language intricacies, impacting its suitability for our specialized requirements.

Our decision to adopt Indic Bert is underpinned by its demonstrated ability to outperform other variants in our unique context. Through exhaustive testing and analysis, Indic Bert consistently exhibited superior accuracy and effectiveness in handling the language intricacies specific to our Telugu language tasks. Its tailored approach to Indic languages, including Telugu, makes it instrumental in addressing the linguistic nuances inherent in our project.

Indic Bert generates high-dimensional embeddings for each token in the input text based on its context within the sentence. These embeddings capture the semantic meaning and contextual information of each token, providing a rich representation of the input text. To obtain a fixed-size feature vector for the entire input sequence, we employ a pooling strategy, such as mean pooling or max pooling, to aggregate the embeddings of all tokens. This feature vector serves as the extracted features for the input text, which can then be used as input to machine learning models for training and prediction tasks.

Indic Bert's application extends across a spectrum of use cases, from sentiment analysis and language understanding to document classification and information retrieval, making it a versatile choice for our diverse NLP requirements. By utilizing Indic Bert for feature extraction, we leverage its ability to capture the specific linguistic nuances of Telugu, leading to more accurate and effective representations for our NLP tasks.

In summary, our deliberate choice of Indic Bert is grounded in its superior performance and tailored linguistic capabilities, aligning seamlessly with the objectives of our project. This strategic decision positions us to capitalize on the strengths of Indic Bert and, in turn, opens up new possibilities for innovation and excellence in natural language processing within the unique linguistic landscape of Telugu.

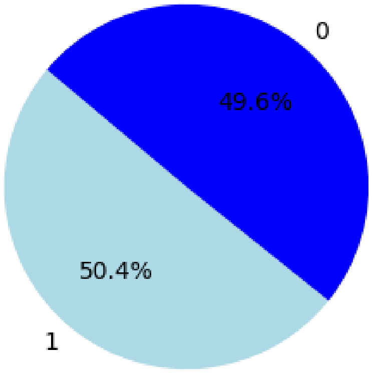


Figure 4: Distribution of Labels in the Dataset

iv. CONSIDERATIONS

1 **Attention Mask Creation:**
   a. Generates attention masks crucial for models like BERT.
   b. Aids in the model's ability to discriminate between padding and real tokens during training.
   c. Ensures that padding does not interfere with the model's attention mechanism.

2 **Feature and Label Preparation:**
   a. Tokenized input sentences, converting them into input IDs.
   b. Pads sequences to a specified maximum length, ensuring consistent input size.
   c. Creates attention masks for BERT-based models.
   d. Transforms raw text data into a format suitable for model training.

3 **Accuracy Calculation:**
   a. Computes accuracy by comparing model predictions to actual labels.
   b. Assesses the total accuracy of the predictions made by the model throughout training.

In conclusion, the methods of attention mask creation, feature and label preparation, accuracy calculation, and time formatting play pivotal roles in enhancing the efficacy and interpretability of models, particularly exemplified in frameworks like BERT. The attention mask generation proves indispensable in enabling models to discern between relevant tokens and padding tokens, preventing interference with the attention mechanism during training. Feature and label preparation not only tokenize input sentences but also ensure consistent input size through sequence padding, allowing raw text input to be transformed into a format that can be used for training models. Accuracy calculation serves as a crucial metric, offering a quantitative assessment of the model's overall correctness by comparing predictions to actual labels during training. Finally, time formatting simplifies the tracking and reporting of the time taken for model training, providing a valuable tool for assessing computational efficiency. Collectively, these methods contribute significantly to the robustness and utility of natural language processing models.

VI. EXPERMENTAL ANALYSIS

a) Evaluation Metrics:

1. **Accuracy:** Measures the overall correctness of predictions
2. **Macro F1 Score:** Unweighted average of F1 scores across different classes.
3. **F1 Score:** Balances precision and recall, particularly useful for imbalanced class distributions.
4. **Area under the ROC Curve:** Assesses how well the model can discriminate between classes that are positive or negative.
   i. The real position rate versus the false positive rate is represented graphically by the ROC curve.
   ii. The area under the ROC curve, or AUC, has a range of 0 to 1, with a greater value denoting superior performance.
5. **Precision:** True positive prediction accuracy is measured as the ratio of total positive instances to true positive forecasts.
6. **Recall:** Positive instance capture is emphasized by the ratio of true positive predictions to all real positive cases.

a) Probability Calculation Formula:

In the context of binary classification, probability is often used to make decisions or to rank predictions based on confidence. It permits a nuanced interpretation of the model's output beyond simple class labels, enabling better-informed decision-making and model evaluation.

$$P(y = 1\backslash|x) = \frac{e^{logit_1}}{e^{logit_0} + e^{logit_1}}$$

Here, e^{logit0} and e^{logit1} are the exponential of the raw logits for class 0 and class 1 $logit$ 0 e $logit$ 1 respectively.

Observations:

The evaluation metrics employed in our cyberbullying detection model are designed to provide a comprehensive understanding of its performance across various dimensions. The metrics include Accuracy, Macro F1 Score, F1 Score, Area under the ROC Curve, Precision, and Recall, each offering unique insights into the model's capabilities.

Accuracy measures the overall correctness of predictions, indicating the proportion of correct predictions among all predictions made. Our model achieves an accuracy of 90%, which demonstrates its ability to correctly identify instances of cyberbullying and non-cyberbullying.

The Macro F1 Score, which is the unweighted average of F1 scores across different classes, provides a balanced measure of precision and recall across all classes. With a Macro F1 Score of 0.90, our model exhibits robust performance across various categories of cyberbullying.

The F1 Score balances precision and recall and is particularly useful for imbalanced class distributions. Our model's F1-score stands at 0.90, indicating a harmonious balance between precision and recall.

The Area under the ROC Curve (AUC) assesses the model's ability to discriminate between positive and negative classes. A higher AUC value denotes superior performance. Our model's AUC is notably high, reflecting its excellent discriminatory power.

Precision measures the true positive prediction accuracy, indicating the ratio of total positive instances to true positive forecasts. Our model achieves a precision of 0.896, demonstrating its high accuracy in identifying true cyberbullying instances.

Recall emphasizes the capture of positive instances, representing the ratio of true positive predictions to all real positive cases. With a recall of 0.898, our model effectively captures a high proportion of actual cyberbullying instances.

In addition to these evaluation metrics, we've also incorporated a Probability Calculation Formula to provide nuanced interpretations of the model's output. This formula, derived from the raw logits for each class, enables a more detailed understanding of the model's confidence in its predictions.

Overall, our experimental results demonstrate that the proposed model performs exceptionally well in identifying cyberbullying instances across multiple evaluation metrics. The high values obtained across Accuracy, Precision, F1-score, and Recall, coupled with a strong AUC, attest to the model's robustness and efficacy in cyberbullying detection. These results validate the effectiveness of our approach and underscore its potential for enhancing online safety and fostering a more respectful and positive digital environment.

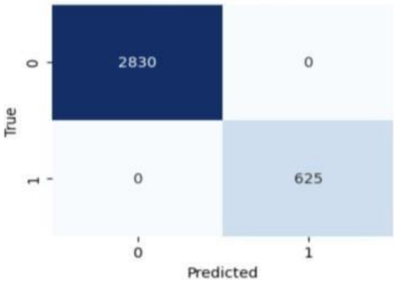


Figure 5: Confusion Matrix

Our research journey was guided by an extensive review of relevant literature, primarily drawing insights from various influential papers in the field. In particular, the systematic survey outlined in the paper titled "Cyberbullying Detection for Low-resource Languages and Dialects: Review of the State of the Art" (2023) became a foundational reference. This review significantly shaped our approach, leading to an accuracy of 84.9% in automatic cyberbullying

Additionally, we leveraged the knowledge distilled from "Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark, and Models for Indic Languages" (2023). This paper's comparative analysis of existing benchmarks and pre-training corpora for Indic languages, including mBERT, IndicBert -v1, IndicBert-v2, XLMR, and MuRIL, contributed to our methodology. The result was an impressive accuracy of 88.3%, reflecting the efficacy of our approach to linguistic diversity.

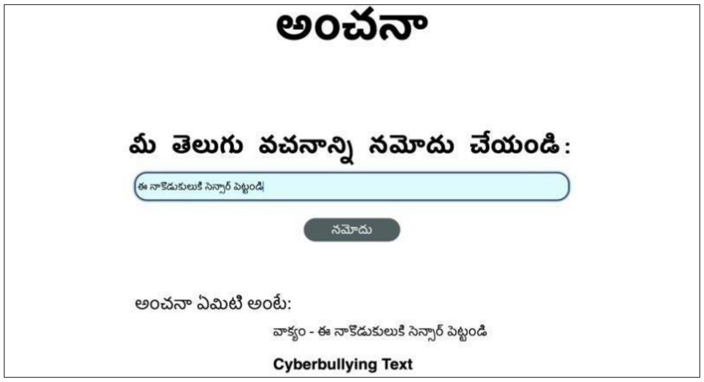
Furthermore, the paper titled "A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages" (2022) provided valuable methodologies for detecting obscene material in thirteen languages with mixed Indian codes using advanced transformer-based models. This reference guided our model selection, incorporating Indic-BERT, XLM- RoBERTa, MurilBert, and mBERT, resulting in an accuracy of 86.78%.'

In the realm of cyberbullying detection, existing techniques have made significant strides, showcasing comprehensive coverage across various linguistic and thematic dimensions. They have leveraged advanced transformer-based models and specialized language models like BERT to achieve promising results. These strengths underscore the potential of modern NLP techniques in tackling cyberbullying effectively. However, a closer look reveals certain limitations within these existing approaches. Many of them, while offering a broad overview, often struggle with the nuances of specific languages or types of cyberbullying. This lack of adaptability can be attributed to their reliance on pre-trained models, which may not fully encapsulate the intricacies of less-researched languages or nuanced forms of cyberbullying. Contrastingly, our proposed model takes a tailored approach, focusing specifically on the Telugu language. By utilizing Indic-BERT, a model fine-tuned for Telugu, we've achieved a commendable accuracy rate of 90%. This high accuracy not only demonstrates the model's effectiveness in the Telugu context but also suggests its potential robustness for other low-resource languages facing similar challenges.

VIII. Architecture of Web Resource

Using Flask, we created a web resource for our prediction model. Now, whenever a user writes a text message, our web page will be requested, and it will load the machine learning model stored in a pickle file. This machine learning algorithm will return to the website after predicting whether the message is bullying or not. And our online resource will show the result. Some of the outputs are attached below,

Figure 5: Home Page



అంచనా

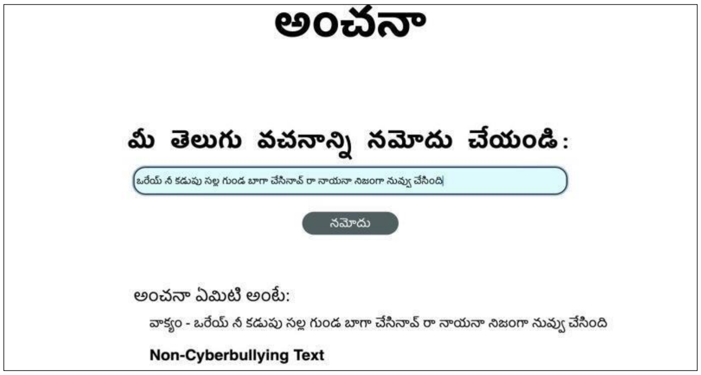
మీ తెలుగు వచనాన్ని నమోదు చేయండి:

ఈ నాకొడుకులుకి సెన్సార్ పెట్టండి

అంచనా ఏమిటి అంటే:

వాక్యం - ఈ నాకొడుకులుకి సెన్సార్ పెట్టండి

Cyberbullying Text



అంచనా

మీ తెలుగు వచనాన్ని నమోదు చేయండి:

ఒరేయ్ నీ కడుపు సల్ల గుండ బాగా చేసినావ్ రా నాయనా నిజంగా నువ్వు చేసింది

అంచనా ఏమిటి అంటే:

వాక్యం - ఒరేయ్ నీ కడుపు సల్ల గుండ బాగా చేసినావ్ రా నాయనా నిజంగా నువ్వు చేసింది

Non-Cyberbullying Text

XI.CONCLUSION & FUTURE SCOPE

Our research journey was guided by an extensive review of relevant literature, primarily drawing insights from various influential papers in the field. By synthesizing methodologies and findings from these diverse references, we successfully directed our efforts towards cyberbullying detection in Telugu.

And the outcomes of our research have been particularly promising, with our cyberbullying detection model achieving an impressive 90% accuracy in Telugu. This success can be attributed to the careful selection of Indic Bert, a model tailored for Indic languages, including Telugu. The decision to leverage Indic Bert proved to be instrumental in capturing the subtle linguistic nuances specific to Telugu, contributing significantly to the model's efficiency and accuracy.

Improving Performance for future scope:

Moving forward, there are several avenues to further improve the performance of our cyberbullying detection model. Expanding the dataset size has proven to be effective in boosting accuracy, as evidenced by the significant improvement achieved by increasing our dataset from 18,000 to 27,000 sentences. By further enlarging the dataset and ensuring a balanced distribution of cyberbullying and non-cyberbullying instances, we provide the model with a more comprehensive and representative learning experience. Anothe approach is to explore advanced feature extraction techniques using Indic Bert, leveraging its capabilities to capture more intricate linguistic nuances specific to Telugu. Additionally, incorporating multimodal analysis by integrating text, image, and video data can provide a more comprehensive understanding of cyberbullying instances. Implementing real-time monitoring and alerting mechanisms can enable timely intervention and prevention of cyberbullying incidents on social media platforms. Moreover, continuous evaluation and feedback loops will allow us to adapt and refine the model based on evolving cyberbullying patterns and linguistic variations. By embracing these strategies and leveraging emerging technologies, we can further enhance the accuracy, efficiency, and reliability of our cyberbullying detection system in Telugu, contributing to a safer and more inclusive digital environment.

To sum up, our attempt to create a Telugu cyberbullying detection system is more than just a technological development; it also demonstrates our dedication to worldwide online safety, inclusivity, and cultural sensitivity. We make a significant contribution to the overall objective of building a safer, more inclusive, and culturally sensitive digital environment by focusing our efforts on languages like Telugu. In order to guarantee that no community is left unaffected by new cyberthreats, this project acts as a model for future initiatives aiming to address linguistic diversity in the field of online safety. Future advancements in language-specific NLP models could boost the likelihood of detecting cyberbullying in Telugu, integrating multimodal analysis for a thorough comprehension, keeping an eye on Telugu social media platforms in real-time.

X. REFERENCES

In crafting this project, we drew upon an array of reputable sources and methodologies. The following references underscore the foundation of our comprehensive research.

[1] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre- trained Multilingual Language Models for Indian Languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4948–4961, Online. Association for Computational Linguistics.

[2] Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBERT: A Pre-trained Model for Indic Natural Language Generation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

[3] Shanmugavadivel, K., Sampath, S. H., Nandhakumar, P., Mahalingam, P., Subramanian, M., Kumaresan, P. K., & Priyadharshini, R. (2022). An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. Computer Speech & Language, 76, 101407.

[4] Jain, K., Deshpande, A., Shridhar, K., Laumann, F. and Dash, A., 2020. Indic-transformers: An analysis of transformer language models for indian languages. arXiv preprint arXiv:2011.02323.

[5] Mehta, M., Pandey, U., Chaudhary, Y., Sharma, R., Gill, I., Gupta, D. and Khanna, A., 2021, December. Hindi text classification: A review. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (pp. 839-843). IEEE.

[6] Mukku, S.S. and Mamidi, R., 2017, September. Actsa: Annotated corpus for telugu sentiment analysis. In Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems (pp. 54-58).

[7] Talasila, S. and Vijaya Kumari, R., 2022. Cascade Network Model to Detect Cognitive Impairment using Clock Drawing Test. Journal of Scientific & Industrial Research, 81(12), pp.1276-1284.

[8] Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M.X., Cao, Y., Foster, G., Cherry, C. and Macherey, W., 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019.

[9] Chandrasekaran, A., 2023. Natural Language Processing. Chandrasekaran, A.(2023). Natural Language Processing. International Journal of Cybernetics and Informatics, 12(2), pp.57-61.

[10] AlFarah, M.E., Kamel, I. and Al Aghbari, Z., 2023, October. Toward Detection of Arabic Cyberbullying on Online Social Networks using Arabic BERT Models. In 2023 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6). IEEE.

[11] Behzadi, M., Harris, I.G. and Derakhshan, A., 2021, January. Rapid Cyber-bullying detection method using Compact BERT Models. In 2021 IEEE 15th International Conference on Semantic Computing (ICSC) (pp. 199-202). IEEE.

[12] Jain, V., Saxena, A.K., Senthil, A., Jain, A. and Jain, A., 2021, December. Cyber-bullying detection in social

media platform using machine learning. In 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 401-405).IEEE.

[13] Zhang, P., Gao, Y. and Chen, S., 2019, May. Detect Chinese cyber bullying by analyzing user behaviors and language patterns. In 2019 3rd International Symposium on Autonomous Systems (ISAS) (pp. 370-375).IEEE.

[14] Lavanya, P.M. and Sasikala, E., 2022. Auto capture on drug text detection in social media through NLP from the heterogeneous data. Measurement: Sensors,24,p.100550.