Multilingual Counter speech Generation

A PROJECT REPORT

Submitted by,

Mr. Avyukth Potnuru-	**20211CAI0123**
Mr. Ayush Samuel Ajith-	**20211CAI0092**
Mr. Naheel N Akhtar-	**20211CAI0142**
Mr. Hasan Raza B A -	**20211CAI0091**

Under the guidance of,

Mr. Likhith S R
Assistant Professor

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY

BENGALURU

JANUARY 2025

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the Project report **"Multilingual Counter speech Generation"** being submitted by "Avyukth Potnuru, Ayush Samuel Ajith, Naheel N Akhtar, Hasan Raza B A" bearing roll number(s) "20211CAI0123, 20211CAI0091, 20211CAI0142, 20211CAI0092" in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out under my supervision.

16/1/25

Mr. Likhith S R
Assistant Professor
School of CSE
Presidency University

16/01/2025

Dr.Zafar Ali Khan N
~~Associate~~ Professor & HoD
School of CSE&IS
Presidency University

Dr. L. SHAKKEERA
Associate Dean
School of CSE
Presidency University

Dr. MYDHILI NAIR
Associate Dean
School of CSE
Presidency University

Dr. SAMEERUDDIN KHAN
Pro-Vc School of Engineering
Dean -School of CSE&IS
Presidency University

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **Multilingual Counter speech Generation** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **Mr. Likhith S R, School of Computer Science Engineering & Information Science, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

Name of the Student	Roll Number	Signature
Avyukth Potnuru	20211CAI0123	Avyukth.P
Ayush Samuel Ajith	20211CAI0091	
Hasan Raza B A	20211CAI0092	Hasan
Naheel N Akhtar	20211CAI0142	

ABSTRACT

In this project we aim to focus on one such solution - Multilingual counter-speech generation using Generative AI, which allows for state-of-the-art natural language processing techniques to successfully detect hate speech, analyse its sentiment and toxicity and generate culturally relevant, informative, and appropriate counter narratives that can curb the mindset of people. The system is designed to be able to process user input, detect the language of the hate speech to ensure global coverage, and classify it into target groups such as religion, caste, colour, or gender. Techniques such as TF-IDF vectorization and cosine similarity allow us to identify relevant examples from curated datasets to inform the analysis. The sentiment and toxicity scores allow for the system to process the severity and intention behind the inputted hate speech. To easy the user's interaction with the model, we have included a user interface created using the Gradio library. Through rigorous evaluation and validation techniques, the project emphasizes cultural sensitivity, linguistic accuracy, and the ethical implications of AI-driven counter-speech.

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC, School of Engineering and Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Deans **Dr. Shakkeera L and Dr. Mydhili Nair,** School of Computer Science Engineering & Information Science, Presidency University, and **Dr. Zafar Ali Khan N**, Head of the Department, School of Computer Science Engineering & Information Science, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Mr. Likhith S R, Assistant Professor** and Reviewer **Mr. Pavan Kumar S P, Assistant Professor**, School of Computer Science Engineering, Presidency University for his inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the PIP2001 Capstone Project Coordinators **Dr. Sampath A K, Dr. Abdul Khadar A and Mr. Md Zia Ur Rahman,** department Project Coordinators **Dr. Afroz Pasha, Assistant Professor** and Git hub coordinator **Mr. Muthuraju V.**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

Avyukth Potnuru (1)
Ayush Samuel Ajith (2)
Naheel N Akhtar (3)
Hasan Raza B A (4)

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

Hate speech refers to any form of communication that gives rise to hatred, violence, or discrimination against individuals or groups based on characteristics such as race, ethnicity, nationality, religion, gender, sexual orientation, or disability etc. The evolution of the internet and social media has drastically transformed the usage of hate speech, enabling it to spread more rapidly and widely than before. Historically, hate speech was often limited to physical gatherings or printed materials, but online platforms allow this to global audiences who are using their services/platforms.

One of the significant factors contributing to the rise of online hate speech is the safety net provided by social media and other online forums. This often gives confidence to individuals to express hateful sentiments they might otherwise refrain from sharing in public. This also gives rise to lack of accountability in the individuals using hate speech, as they feel safe from the consequences of their actions.

Legal frameworks surrounding online hate speech vary significantly across countries. Some nations impose strict regulations and penalties for hate speech, while others prioritize the protection of free speech, complicating the ability to effectively address this issue. This legal ambiguity presents challenges for both online platforms and policymakers, as they navigate the delicate balance between fostering free expression and combating harmful rhetoric.

The impact of online hate speech in our society is profound, because it can lead to real-world consequences such as violence, discrimination, and mental health issues for the targeted individuals. It also creates a hostile online environment where people are able to come together and bully people who don’t agree with them. Recognizing the severity of the problem, researchers and technology developers are turning to machine learning and natural language processing techniques to detect and mitigate hate speech online. However, these efforts face significant challenges, particularly in balancing the necessity of combating hate speech with the fundamental right to free expression.

In response to the growing concern over online hate speech, various organizations, civil society groups, and social media platforms have initiated campaigns to educate users, implement reporting mechanisms, and enforce policy changes. Grassroots movements have also emerged to counteract hate speech with positive messaging, advocating for inclusivity and understanding. Addressing online hate speech is a complex challenge that requires ongoing research, dialogue, and collaborative efforts among stakeholders to effectively tackle its root causes and mitigate its detrimental effects on society.

CHAPTER-2

LITERATURE SURVEY

Table 2.1: Study of Tools and Methodologies

References No	Year	Study of Tools/Technology	Overall Accuracy	Dataset
[1]	2024	CoARL method, which utilizes multi-task instruction tuning, Low-Rank Adapters (LoRA) for intent adaptation, and Proximal Policy Optimization (PPO) for fine-tuning.	CoARL outperforms baseline models in generating effective, high-quality counter speech aligned with specific intents.	IntentCONANv2
[2]	2024	A survey of Natural Language Processing (NLP) techniques used to develop counter speech systems against online hate speech.	The excerpt does not specify any accuracy metrics related to the counter speech systems.	Does not mention any specific datasets used in the study.
[3]	2024	Multilingual large language models (LLMs) using external knowledge from ConceptNet and Wikipedia through language and task adapters.	Results indicate that integrating external knowledge significantly improves model performance in sentiment analysis and named entity recognition for various LRLs.	The focus is more on leveraging external knowledge rather than detailing a specific dataset.

[4]	2024	Prompting with instructions, selecting from multiple generated responses, fine-tuning LLMs, and using reinforcement learning to improve counter speech generation.	The excerpt does not provide specific accuracy metrics for the methods evaluated.	The paper does not mention any specific datasets used for the counter speech generation methods discussed.
[5]	2024	mT5 multilingual model for counter-narrative (CN) generation.	The results indicate that post-edited data significantly improves CN generation, particularly for Spanish; specific accuracy metrics are not provided in the excerpt.	The CONAN-EUS dataset
[6]	2023	RAUCG (Retrieval-Augmented Unsupervised Counter Narrative Generation) and utilizes models like GPT-2, DiabloGPT-ft, and GPS.	The excerpt does not provide specific accuracy metrics for the performance of the RAUCG framework..	Web scraped from social media platforms such as YouTube, Reddit, and Gab. Additionally, experiments were conducted on the Multitarget CONAN dataset.
[7]	2024	mBERT, DistilBERT, IndicBERT, and others, for hate speech detection in the Telugu language.	mBERT achieved an accuracy of 98.2% in classifying hate speech and non-hate speech.	A monolingual dataset of 38,000 annotated tweets in the Telugu language was created for the study.

[8]	2024	SCoRe, method that utilizes reinforcement learning (RL) to enable language models to self-correct their mistakes over multiple interactions.	SCoRe achieves a 15.6% improvement in math tasks and a 9.1% improvement in coding tasks compared to traditional fine-tuning methods.	The paper does not specify a particular dataset used for the experiments.
[9]	2024	Counter speech (CS) in low-resource languages, specifically Bengali and Hindi, using transformer-based models such as GPT-2, mT5, BLOOM, and ChatGPT.	It highlights the superiority of monolingual models over joint or synthetic transfer models, particularly for related languages.	A dataset of 5,062 abusive speech and counter speech pairs was created to support the research.
[10]	2024	A Spanish corpus for generating counter speech (CN) using language models like GPT-3 and GPT-4.	GPT-4 performs significantly better than GPT-3.	MultiTarget CONAN dataset in English, which covers various hate targets, which serves as the basis for the Spanish corpus.
[11]	2024	The study evaluates various large language models (LLMs) for zero-shot counter speech generation, including GPT-2, DialoGPT, FlanT5, and ChatGPT.	GPT-2 outperforms the other models in terms of generation quality, while GPT-2 is noted for its superior quality in counter speech specifically.	The datasets used in the study include CONAN (3,864 pairs), Reddit (14,223 pairs), Gab (41,580 pairs), and CONAN-MT (5,003 pairs), along with additional counter speech and counterargument dat asets.

[12]	2022	DiabloGPT using a three-stage pipeline called Generate Prune Select (GPS). It employs the GEDI framework to manage positive and negative attributes in responses.	The excerpt does not provide specific accuracy metrics for the model's performance but mentions that multi-attribute models perform best across all datasets.	Examples of hate speech and counter speech collected from Reddit and Gab, comprising thousands of instances written by crowd workers or expert NGOs.
[13]	2022	Pre-trained language models for generating automatic counter-narratives (CNs) against online hate speech, comparing models such as BERT, GPT-2, DialoGPT, BART, and T5.	Training on semantically related hate speech targets improves model generalization to unseen targets.	Fanton et al. (2021) dataset, which comprises 5,003 pairs of hate speech and corresponding counter-narratives.
[14]	2021	Seq2Seq with GRU layers, Maximum Mutual Information (MMI), SpaceFusion for response diversity, and BART for sequence generation.	The study indicates that the methods significantly enhance the relevance and diversity of generated counter speech, leading to improved effectiveness compared to traditional approaches.	Reddit and Gab, focusing on user-generated hate speech and counter speech content, while noting the limitations of YouTube data for effective training.

[15]	2021	The proposed approach utilizes the GPT-2 model, specifically fine-tuned with external knowledge. It also incorporates knowledge retrieval from sources such as Newsroom and WikiText-103.	GPT-2_KN is indicated to outperform baseline models like GPT-2 and Candela, suggesting a significant improvement in performance.	The dataset used for this study is the CONAN dataset, which consists of 6,645 hate speech-counter narrative (HS-CN) pairs.
[16]	2024	Large language models (LLMs) for generating counter-narratives (CNs) against hate speech in Spanish, focusing on zero-shot (ZS) generation It tests three approaches: an ensemble model, a ZS model using the top-performing LLM, and a fine-tuned version. The evaluation utilizes JudgeLM to select the most effective CNs.	The fine-tuned model demonstrates superior performance in both automatic and manual evaluations, although manual annotations reveal discrepancies in informativeness.	CONAN-MT-SP dataset.

The paper introduces IntentCONANv2, a dataset with 13,952 counter speech examples across four intents: positive, informative, questioning, and denouncing. It improves counter speech quality and distribution. [1] The proposed method, CoARL, generates intent-specific, non-toxic counter speech using three steps: multi-task instruction tuning, Low-Rank Adapters (LoRA) for intent adaptation, and Proximal Policy Optimization (PPO) for fine-tuning. CoARL outperforms baseline models in generating effective, high-quality counter speech aligned with specific intents.

The paper from the NAACL 2024 findings provides a comprehensive survey and guide on using Natural Language Processing (NLP) to develop counter speech systems against online hate. Counter speech is framed in an effective way to combat hate speech while preserving freedom of speech. The paper reviews the challenges of collecting, classifying, and automatically generating counter speech, highlighting its significance in reducing both online and offline violence. [2] The authors discuss various strategies and best practices from existing NLP research, while identifying open challenges like addressing language, cultural differences and other hateful sentiments. They propose a structured guide for task design, data collection, and evaluation, making it accessible to both newcomers and experts in the field. Finally, the paper emphasizes the ethical and societal responsibilities associated with counter speech, especially in real-world applications.

The paper investigates enhancing multilingual large language models (LLMs) for low-resource languages (LRLs) by integrating external knowledge, specifically from linguistic ontologies like ConceptNet and Wikipedia, through language and task adapters. The study focuses on improving performance in sentiment analysis (SA) and named entity recognition (NER) for LRLs, including Maltese, Uyghur, Tibetan, and others. [3] The research utilizes adapters to inject multilingual graph knowledge, comparing different training objectives like masked language modeling (MLM) and targeted masking. Results show that integrating external knowledge improves model performance in many LRLs, particularly for sentiment analysis. Adapter Fusion combining Wikipedia and ConceptNet data shows promise but requires further refinement. This research highlights the potential of external knowledge for LRLs and emphasizes the need for individualized approaches for each language.

The paper explores methods to generate counter speech for hate speech using large language models (LLMs), with a focus on guiding these models to create responses that achieve specific conversation outcomes, such as reducing incivility or promoting constructive engagement. It presents and evaluates four approaches: prompting with instructions, selecting from multiple generated responses, fine-tuning LLMs, and using reinforcement learning to align text generation with desired outcomes. The study finds that these methods improve the likelihood of generating counter speech that fosters healthier online conversations. [4]

The paper presents CONAN-EUS, a new parallel dataset for counter-narrative (CN) generation in Basque and Spanish, developed using machine translation (MT) and professional post-editing based on the English CONAN dataset. The study focuses on generating counter-narratives to mitigate online hate speech using the mT5 multilingual model, comparing the performance of machine-translated data and post-edited data. [5] The results demonstrate that post-edited data significantly improves CN generation, especially for Spanish, while multilingual data augmentation benefits structurally similar languages like English and Spanish but is less effective for Basque, a language isolate. The study highlights the importance of post-edited training data and multilingual research for CN generation.

This study presents RAUCG (Retrieval-Augmented Unsupervised Counter Narrative Generation), which focuses on generating counter speech (CN) for hate speech (HS) using models like GPT-2, DiabloGPT-ft, and GPS. Datasets for CN generation include 6,898 CNs and 7,026 non-CNs scraped from social media platforms like YouTube, Reddit, and Gab. Experiments were conducted on the MultitargetCONAN dataset, consisting of 5,003 HS-CN pairs. [6] The main challenge involves retrieving relevant "counter-knowledge" that not only opposes HS but also maintains stance consistency and semantic overlap. To address this, a customized retrieval method was developed, integrating these metrics. Another challenge is mapping counter-knowledge to CNs in the absence of expert-authored data, which was solved using energy-based decoding with three constraint functions to ensure the generated CNs counter the HS, retain fluency, and incorporate the retrieved knowledge. The RAUCG framework includes two key components: the Counter Knowledge Retriever (CKR) and the Counter Narrative Generator (CNG), working together to retrieve counter-knowledge and generate unsupervised CNs.

The paper focuses on detecting hate speech in the Telugu language using deep learning transformer models. Given the challenges of low-resource languages like Telugu, the authors created a monolingual dataset of 38,000 annotated tweets. They experimented with various fine-tuned deep learning models, including mBERT, DistilBERT, IndicBERT, and others, to classify hate speech and non-hate speech. mBERT emerged as the best-performing model with an accuracy of 98.2%. The study emphasizes the importance of creating language-specific datasets and models for better hate speech detection and proposes a deployment model for social media platforms. [7]

This paper introduces SCoRe, a method that helps language models correct their own mistakes using reinforcement learning (RL). Instead of relying on external feedback or fine-tuning, SCoRe trains models to self-correct over multiple turns, improving their responses. The approach shows significant performance gains, with a 15.6% improvement in math and 9.1% in coding tasks. SCoRe outperforms traditional fine-tuning, proving more effective for complex reasoning and coding problems. [8] The study further demonstrates through ablation and scaling experiments that RL-based self-correction is more effective than static approaches like fine-tuning, particularly in complex reasoning and coding tasks. SCoRe thus represents a breakthrough in teaching LLMs to improve their problem-solving abilities through intrinsic self-correction.

The research paper focuses on generating counter speech (CS) in low-resource languages, specifically Bengali and Hindi, to combat online abuse. The authors created a dataset of 5,062 abusive speech and counter speech pairs. They experimented with various transformer-based models, including GPT-2, mT5, BLOOM, and ChatGPT. The study found that monolingual models outperformed joint or synthetic transfer models, particularly when languages shared a family (e.g., Bengali and Hindi). [9] Challenges included generating diverse counter speech and ensuring models produced non-abusive, coherent responses.

This study introduces CONAN-MT-SP, a Spanish corpus for generating counter speech (CN) using models like GPT-3 and GPT-4. The dataset is based on the MultiTarget CONAN dataset, which includes 5,003 pairs of hate speech (HS) and counter speech (CN) in English, covering various hate targets such as race, religion, nationality, sexual orientation, disability, and gender. [10] A key challenge identified is the generation of "default CNs," where the structure of the CN remains generic and lacks topic-specific relevance, reducing its effectiveness by not adding meaningful information. The study evaluates the performance of GPT-4 against GPT-3, with findings suggesting that GPT-4 offers improvements in terms of quality and consistency in CN generation. However, minor issues such as lexical, orthographic, and grammatical errors were noted, which can affect the coherence and comprehension of the generated CNs.

This study explores zero-shot counter speech generation using large language models (LLMs) across four datasets: CONAN (3,864 pairs), Reddit (14,223 pairs), Gab (41,580 pairs), and CONAN-MT (5,003 pairs), along with additional counter speech and counterargument datasets. Various LLMs, including GPT-2, DialoGPT, FlanT5, and ChatGPT, are evaluated using manual, frequency-based, and cluster-centered prompting strategies. [11] Results show that ChatGPT outperforms other models in generation quality, while GPT-2 excels in counter speech quality. The findings highlight the potential of LLMs for effective counter speech generation and the need for improved prompting strategies.

This study introduces DiabloGPT, a model fine-tuned on hate and counter speech data, using a three-stage pipeline called Generate Prune Select (GPS) to create diverse and appropriate responses. [12] The model leverages the GEDI framework to control positive and negative attributes in the responses. The datasets, from Reddit and Gab, include thousands of hate speech examples with counter speech written by either crowdworkers or expert NGOs. Performance is evaluated using language generation metrics and fluency checks, but the key challenge remains ensuring counter speech is contextually effective. Results show that controlling for multiple attributes like politeness and emotional tone leads to better counter speech, with multi-attribute models performing best across all datasets.

The paper evaluates the performance of pre-trained language models (LMs) for generating automatic counter-narratives (CNs) to combat online hate speech. The study compares various models, including BERT, GPT-2, DialoGPT, BART, and T5, alongside different decoding mechanisms (e.g., Beam Search, Top-k, Top-p). [13] Autoregressive models with stochastic decoding methods were found to generate more diverse and novel CNs, while deterministic methods like Beam Search produced safer but more generic responses. The researchers fine-tuned LMs on the Fanton et al. (2021) dataset, which contains 5003 pairs of hate speech and counter-narratives. In out-of-target experiments, they observed that training data with semantically related hate speech targets improved generalization to unseen targets. Finally, they explored adding an automatic post-editing step to enhance the quality of generated CNs. Key challenges include model bias, the balance between diversity and safety, and the need for high-quality datasets.

This study examines multiple models for generating counter speech (CN) in response to hate speech (HS). Models include Seq2Seq, which uses GRU layers for encoding and decoding, MMI for incremental learning and robust training, SpaceFusion, which balances response diversity and relevance, and BART, a state-of-the-art transformer model for sequence-to-sequence generation. [14] The models are tested on datasets from Reddit and Gab, and performance is measured based on relevance and diversity, with solutions like Cosine Similarity and human evaluation addressing challenges of staying on-topic. A key issue is the tendency of models to generate safe, generic responses, which GPS counters by prioritizing diversity. Observations highlight that counter speech is more effective on platforms like YouTube, where popular comments are perceived as more credible. However, YouTube data lacks the necessary quality for training models, especially in terms of virality and psychological impact. The study suggests the need for more comprehensive datasets to develop effective counter speech.

The proposed approach fine-tunes GPT-2 models with external knowledge from the CONAN dataset, consisting of 6,645 HS-CN pairs. The GPT-2KN model, combined with retrieved knowledge from sources like Newsroom and WikiText-103, outperforms baselines like GPT-2 and Candela in generating more informative CNs. This method shows that knowledge injection significantly enhances CN generation, improving both in-domain and cross-domain performance, making it effective for hate speech mitigation and adaptable to similar tasks like dialogue generation. [15]

The paper discusses the use of large language models (LLMs) for generating counter-narratives (CNs) against hate speech (HS) in Spanish as part of the RefutES 2024 shared task. The authors explore zero-shot (ZS) generation capabilities of LLMs and the benefits of fine-tuning for this task. The dataset used, CONAN-MT-SP, includes hate speech targeting various demographics, with CNs generated using GPT-4 and human-reviewed translations. The authors test three models: an ensemble model combining multiple LLMs, a ZS model using the top-performing model, and a fine-tuned version of the best model. JudgeLM is used to evaluate the generated CNs, selecting the most effective ones. [16] The fine-tuned model outperforms the others in both automatic and manual evaluations, though the manual annotation highlighted discrepancies in informativeness. Key challenges include balancing efficiency and model performance, while addressing issues such as hallucinations in LLM-generated content.

CHAPTER-3

RESEARCH GAPS OF EXISTING METHODS

3.1 Bias in Counter Narrative Generation

In many of the existing methods available, they often inherit biases from the training data which can lead to outputs that may unintentionally favour certain perspectives or perpetuate stereotypes. This can make them less effective in addressing the diverse needs of affected communities.

3.2 Lack of High-Quality and Diverse Data

The scarcity of diverse datasets limits the ability of models to generalize across various contexts. This can be quite a problem as it can hamper the generation of effective counter narratives involving underrepresented groups or languages.

3.3 Inability to Identify Targeted Groups

Many approaches fail to accurately detect the specific groups or communities targeted by hate speech. This shortcoming reduces the relevance and impact of the counter narratives, as they may not be able to fully comprehend the nuances that affect the groups.

3.4 Limited Understanding of Intent and Psychological Aspects

Hate speech often carries implicit intentions, such as a threat or passive aggressive tone, which are not adequately captured by existing models. This gap hinders the generation of counter narratives that effectively de-escalate tensions or challenge the underlying motives of hateful content.

3.5 Challenges in Addressing Linguistic and Cultural Nuances

With each language, the subtleties and cultural context can vary significantly, which for low-resource languages can complicate the generation of accurate and culturally sensitive counter narratives.

3.6 Inadequate Evaluation Frameworks

Current evaluation methods are often subjective or overly simplistic, which fails to capture the effectiveness, relevance, and impact of the generated counter narratives. There is a pressing need for robust and standardized evaluation techniques to assess their quality comprehensively.

3.7 Scalability and Deployment Challenges

Despite promising results in controlled environments, many methods face significant challenges when scaled for real-world deployment. Issues such as computational resource requirements, latency, and adaptability to diverse online platforms hinder their practical application.

CHAPTER-4

PROPOSED METHODOLOGY

4.1 Data Preparation

- Dataset Selection and Curation:
  - A dataset which includes instances of hate speech and their corresponding targets (e.g., WOMEN, MIGRANTS, POC) was cleaned and loaded. This dataset contains a column labelled HS (hate speech) and corresponding TARGET values which is pivotal to predicting and generating a counter-narrative.
  - This dataset includes HS from four languages – English, Spanish, Italian and Basque, all with additional background information that is used as context for the model.

- Text Preprocessing:
  - All textual data was cleaned by removing punctuation, special characters, and extra whitespaces to standardize the inputs and ensure vectorization is performed smoothly.
  - Lowercasing was applied to ensure case insensitivity during vectorization.

- Language Detection:
  - A language detection tool (polyglot library) was used to identify the language of the hate speech inputs. The language code was mapped to the four supported languages - English, Spanish, Italian and Basque
  - Defaulting to English was implemented for unsupported languages.

4.2 Hate Speech Labelling and Classification

- TF-IDF Vectorization:
  - A TF-IDF vectorizer is used to transform the inputted hate speech and dataset text into vectorized forms for comparison (cosine similarity).

- Cosine Similarity Calculation:
  - Cosine similarity was computed between the inputted hate speech and dataset

- examples to identify the most relevant examples in the dataset.
- Rows with the highest similarity scores were selected (top 5).

- Threshold Analysis:
  - A threshold of 0.7 was applied to filter examples with strong matches to the input. If no strong matches were found, the highest count of the TARGET column from the top 5 rows is selected.

- Target Identification:
  - The identified hate speech was assigned a target category (e.g., JEWS, POC, LGBT+) using the top 5 selected rows.

4.3 Toxicity Scoring

- Sentiment Analysis:
  - Sentiment analysis, using the Hugging Face sentiment-analysis pipeline, was applied to assess the sentiment of the hate speech input, and is then used to find the toxicity value.

- Toxicity Derivation:
  - A custom metric was defined to calculate the toxicity. Negative sentiment scores were directly mapped to toxicity, while positive/neutral sentiment scores were inverted to assign lower toxicity values. This value aids the user in gauging the intention behind the hate speech.

- Score Interpretation:
  - A high toxicity score (close to 1) indicated strongly negative sentiment and high toxicity, while a middling value (close to 0.5) indicates a neutral sentiment and a low score (close to 0) indicates a positive sentiment.

4.4 Counter-Narrative Generation

- Prompt Design:
  - Carefully crafted prompts were designed to generate counter-narratives using OpenAI’s GPT-3.5-turbo for 4 languages – English, Spanish, Italian, and Basque.

- Two modes, one-shot and few-shot, were implemented to gauge the differences in the generated counter narrative:
- One-shot Mode: Used a single example as context for generating a counter-narrative in the detected language.
- Few-shot Mode: Used multiple examples of hate speech and corresponding counter-narratives in the 4 languages to guide the response generation process.

- Multilingual Capability:
  - Counter-narratives were generated in the language of the detected input. The model could support the 4 languages mentioned before, but can be further extended to include many more languages based on use cases.

4.5 Integration and User Interaction

- Gradio Interface:
  - Using the Gradio library, a user interface was designed to easily facilitate interaction with the system.
  - Users input hate speech text, and the system displayed both the counter-narrative and the calculated toxicity score.

- Public Accessibility:
  - The Gradio interface was launched with public URL sharing, enabling remote access to the application for testing and demonstrations.

4.6 Evaluation and Validation

- Model Performance:
  - Outputs were reviewed to ensure linguistic accuracy and cultural sensitivity in the generated counter-narratives.
- Toxicity and Labelling Accuracy:
  - Toxicity scores were compared with human judgments to validate the effectiveness of the sentiment analysis pipeline.
- Feedback Collection:
  - User feedback was collected to assess the quality of counter-narratives and improve the system iteratively.

CHAPTER-5

OBJECTIVES

- The project aims to implement a system to identify hate speech in multiple languages by comparing user input with a predefined dataset of harmful content.

- Using sentiment analysis to assess the toxicity of hate speech and provide an overall toxicity score to determine the severity of the input.

- Calculate cosine similarity to compare the user’s input against known examples of hate speech and identify relevant matches.

- Generate a diverse set of counter-narrative examples for different types of hate speechs targetting various groups (e.g., women, POC, LGBT+).

- Develop a user-friendly interface using Gradio to enable users to easily input hate speech and receive counter-narratives and toxicity scores, mitigating toxicity through counter-speech promotes an environment of understanding, diversity, and inclusion.

CHAPTER-6

SYSTEM DESIGN & IMPLEMENTATION

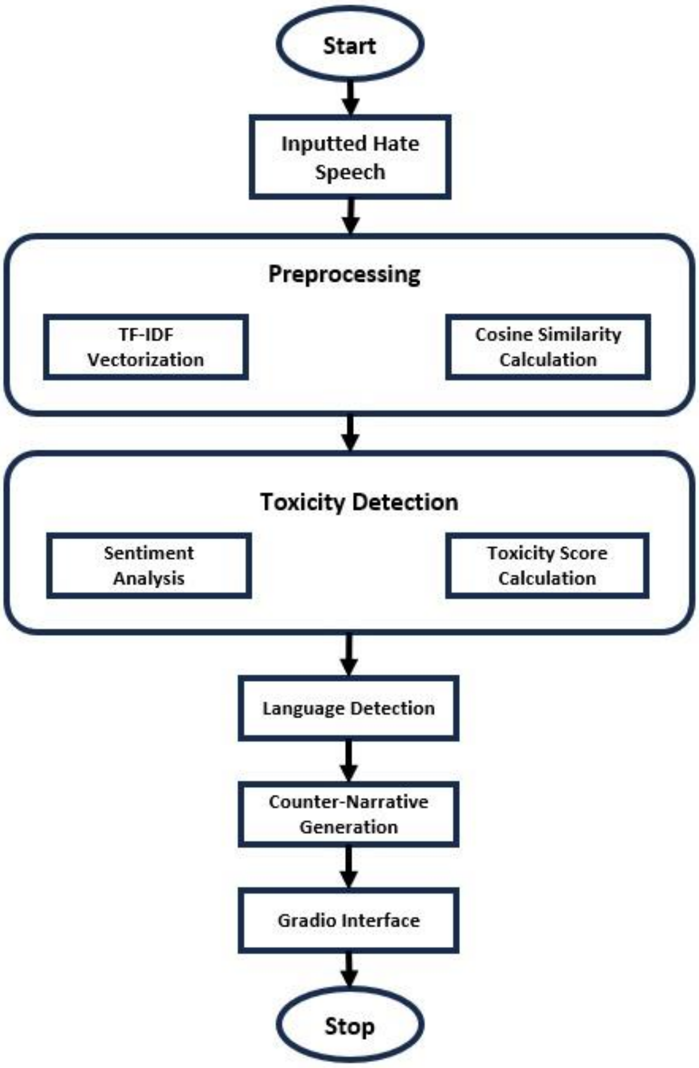


Fig 6.1: Architecture of the model

The figure 6.1 shows us the architecture processes of hate speech step-by-step to generate positive responses. It starts by detecting the language and analyzing the input using text matching and sentiment analysis for toxicity scoring. Based on the language and severity, it uses GPT models to create counter-narratives, either with examples (few-shot) or a single prompt (one-shot). A Gradio interface makes it easy for users to input hate speech and receive real-time responses with toxicity scores.

CHAPTER-7

TIMELINE FOR EXECUTION OF PROJECT

(GANTT CHART)

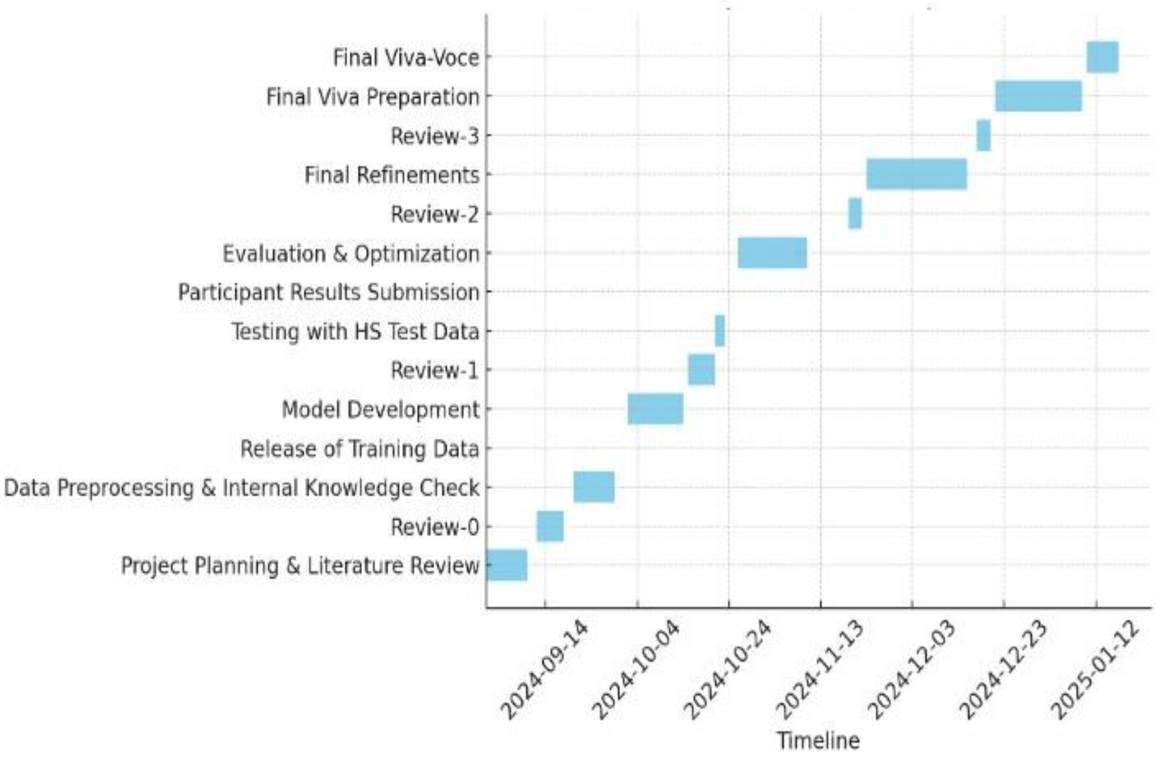


Fig 7.1: Gantt Chart/Project Timeline

The figure 7.1 shows us the timeline of the project in each sequential phase. The timeline represents a structured workflow for a project, beginning with project planning, literature review, and data preprocessing in mid-September 2024. It progresses through training data release, model development, and internal reviews, followed by testing with hate speech (HS) test data and evaluation of results in November-December 2024. Final refinements, reviews, and viva preparation conclude the project by mid-January 2025.

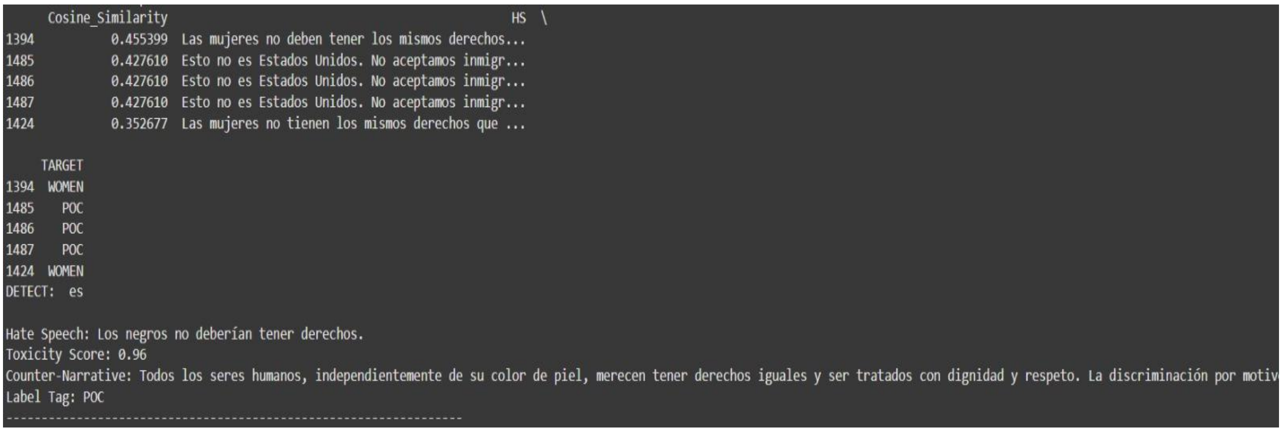
CHAPTER-8

OUTCOMES

- Context-aware responses with external knowledge integration.
- Accurate multilingual output in Basque, English, Italian, and Spanish.
- Ability to handle diverse hate speech across multiple target groups.
- Dynamic retrieval compensates for gaps in internal knowledge.
- Cross-domain adaptability with multi-task instruction tuning.
- Balanced, non-confrontational counter-speech strategies.
- Mitigation of biases using one-shot and few-shot prompting.
- Promotes respectful discourse and combats online hate speech effectively.

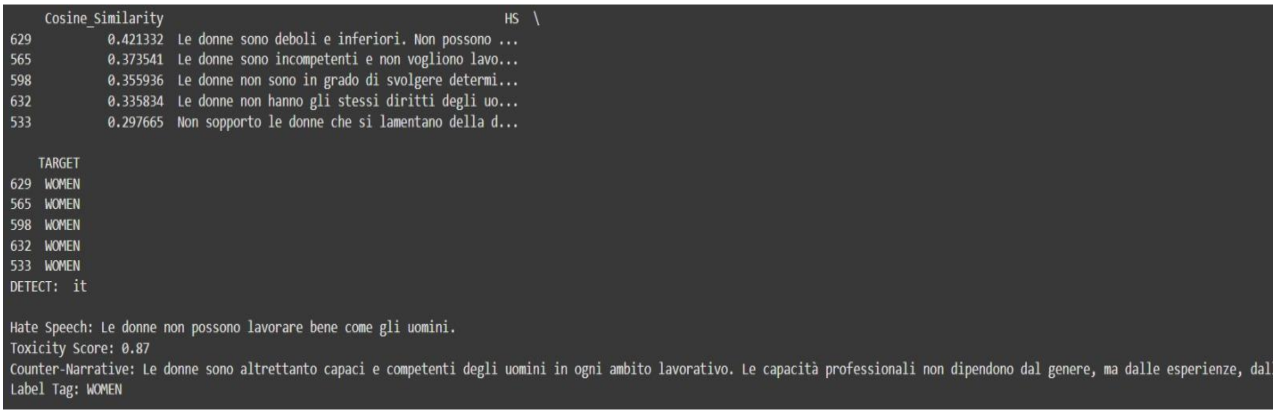
CHAPTER-9

RESULTS AND DISCUSSIONS



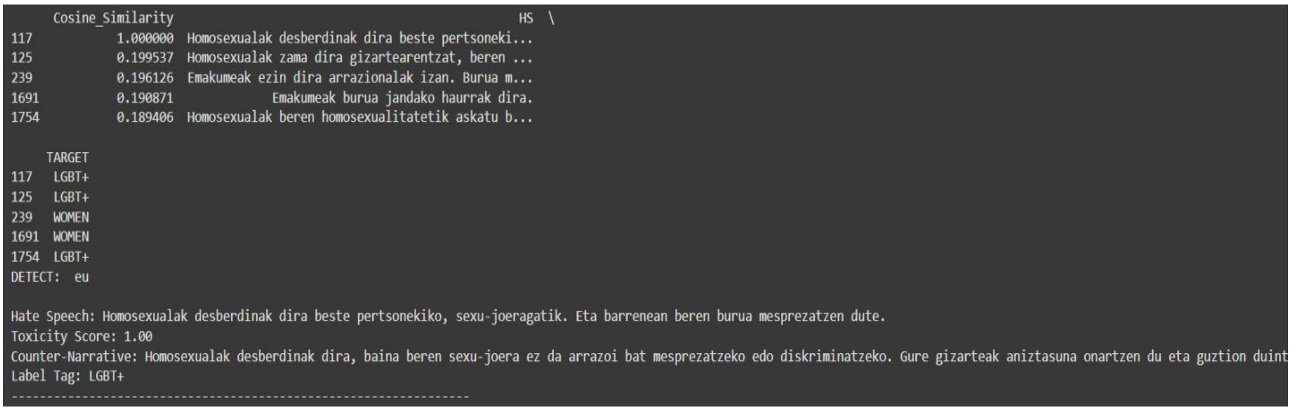
```
     Cosine_Similarity                                            HS  \
1394          0.455399  Las mujeres no deben tener los mismos derechos...
1485          0.427610  Esto no es Estados Unidos. No aceptamos inmigr...
1486          0.427610  Esto no es Estados Unidos. No aceptamos inmigr...
1487          0.427610  Esto no es Estados Unidos. No aceptamos inmigr...
1424          0.352677  Las mujeres no tienen los mismos derechos que ...

    TARGET
1394  WOMEN
1485    POC
1486    POC
1487    POC
1424  WOMEN
DETECT:  es

Hate Speech: Los negros no deberían tener derechos.
Toxicity Score: 0.96
Counter-Narrative: Todos los seres humanos, independientemente de su color de piel, merecen tener derechos iguales y ser tratados con dignidad y respeto. La discriminación por motiv
Label Tag: POC
-----------------------------------------------------------------
```

Fig 9.1,
Displays the top 5 cosine similarity values for generated Spanish counter-narrative



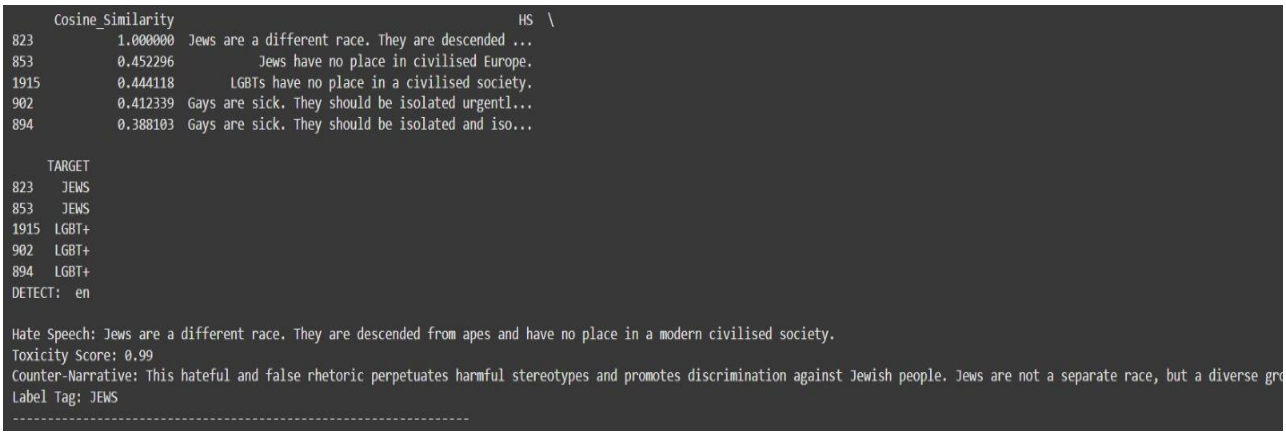
```
     Cosine_Similarity                                            HS  \
629           0.421332  Le donne sono deboli e inferiori. Non possono ...
565           0.373541  Le donne sono incompetenti e non vogliono lavo...
598           0.355936  Le donne non sono in grado di svolgere determi...
632           0.335834  Le donne non hanno gli stessi diritti degli uo...
533           0.297665  Non sopporto le donne che si lamentano della d...

    TARGET
629  WOMEN
565  WOMEN
598  WOMEN
632  WOMEN
533  WOMEN
DETECT:  it

Hate Speech: Le donne non possono lavorare bene come gli uomini.
Toxicity Score: 0.87
Counter-Narrative: Le donne sono altrettanto capaci e competenti degli uomini in ogni ambito lavorativo. Le capacità professionali non dipendono dal genere, ma dalle esperienze, dal
Label Tag: WOMEN
```

Fig 9.2,
Displays the top 5 cosine similarity values for generated Italian counter-narrative



```
      Cosine_Similarity                                          HS  \
117            1.000000  Homosexualak desberdinak dira beste pertsoneki...
125            0.199537  Homosexualak zama dira gizartearentzat, beren ...
239            0.196126  Emakumeak ezin dira arrazionalak izan. Burua m...
1691           0.190871              Emakumeak burua jandako haurrak dira.
1754           0.189406  Homosexualak beren homosexualitatetik askatu b...

     TARGET
117   LGBT+
125   LGBT+
239   WOMEN
1691  WOMEN
1754  LGBT+
DETECT:  eu

Hate Speech: Homosexualak desberdinak dira beste pertsonekiko, sexu-joeragatik. Eta barrenean beren burua mesprezatzen dute.
Toxicity Score: 1.00
Counter-Narrative: Homosexualak desberdinak dira, baina beren sexu-joera ez da arrazoi bat mesprezatzeko edo diskriminatzeko. Gure gizarteak aniztasuna onartzen du eta guztion duint
Label Tag: LGBT+
-----------------------------------------------------------------
```

Fig 9.3,
Displays the top 5 cosine similarity values for generated Basque
counter-narrative



```
      Cosine_Similarity                                          HS  \
823            1.000000  Jews are a different race. They are descended ...
853            0.452296            Jews have no place in civilised Europe.
1915           0.444118        LGBTs have no place in a civilised society.
902            0.412339  Gays are sick. They should be isolated urgentl...
894            0.388103  Gays are sick. They should be isolated and iso...

     TARGET
823    JEWS
853    JEWS
1915  LGBT+
902   LGBT+
894   LGBT+
DETECT:  en

Hate Speech: Jews are a different race. They are descended from apes and have no place in a modern civilised society.
Toxicity Score: 0.99
Counter-Narrative: This hateful and false rhetoric perpetuates harmful stereotypes and promotes discrimination against Jewish people. Jews are not a separate race, but a diverse gr
Label Tag: JEWS
-----------------------------------------------------------------
```

Fig 9.4,
Displays the top 5 cosine similarity values for generated English
counter-narrative

CHAPTER-10

CONCLUSION

This project successfully addresses the need for multilingual counter-speech generation by leveraging advanced AI models, sentiment analysis, and dynamic language mapping techniques. The integration of tools like Hugging Face's pipelines, TF-IDF vectorization, and OpenAI's GPT-3.5-turbo ensures a robust and adaptive framework capable of analyzing and responding to hate speech across various languages and cultural contexts. The system's ability to classify hate speech targets and assess toxicity levels ensures that the counter-narratives are not only relevant but also tailored to the severity and specific audience, enhancing the overall impact and effectiveness of counter-speech in mitigating online hate.

By combining innovative methods such as one-shot and few-shot prompting with user-friendly interfaces like Gradio, the project demonstrates scalability and accessibility for non-technical users and organizations. Additionally, the focus on low-resource languages and contextual similarity in hate speech detection fills critical gaps in existing research, making it a versatile tool for real-world applications. This project sets a strong foundation for future advancements in counter-speech generation and highlights the importance of combining linguistic, cultural, and ethical considerations to combat hate speech in a globally interconnected digital environment.

REFERENCES

[1]. Amey Hengle, Aswini Padhi, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Intent-conditioned and Non-toxic Counter speech Generation using Multi-Task Instruction Tuning with RLAIF. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6716–6733, Mexico City, Mexico. Association for Computational Linguistics.

[2]. Bonaldi, Helena, Yi-ling Chung, Gavin Abercrombie and Marco Guerini. "NLP for Counter speech against Hate: A Survey and How-To Guide." NAACL-HLT (2024).

[3]. Gurgurov, Daniil, Mareike Hartmann and Simon Ostermann. "Adapting Multilingual LLMs to Low-Resource Languages with Knowledge Graphs via Adapters." ArXiv abs/2407.01406 (2024): n. pag.

[4]. Hong, L., Luo, P., Blanco, E., & Song, X. (2024). Outcome-Constrained Large Language Models for Countering Hate Speech. ArXiv, abs/2403.17146.

[5]. Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2132–2141, Torino, Italia. ELRA and ICCL.

[6]. Jiang, S., Tang, W., Chen, X., Tang, R., Wang, H., & Wang, W. (2023). RAUCG: Retrieval-Augmented Unsupervised Counter Narrative Generation for Hate Speech. ArXiv, abs/2310.05650.

[7]. Khanduja, N., Kumar, N., & Chauhan, A. (2024). Telugu Language Hate Speech Detection using Deep Learning Transformer Models: Corpus Generation and Evaluation. Systems and Soft Computing, 200112. ISSN 2772-9419.

[8]. Kumar, Aviral & Zhuang, Vincent & Agarwal, Rishabh & Su, Yi & Co-Reyes, John & Singh, Avi & Baumli, Kate & Iqbal, Shariq & Bishop, Colton & Roelofs, Rebecca & Zhang, Lei & McKinney, Kay & Shrivastava, Disha & Paduraru, Cosmin & Tucker, George & Precup, Doina & Behbahani, Feryal & Faust, Aleksandra. (2024). Training Language Models to Self-Correct via Reinforcement Learning. 10.48550/arXiv.2409.12917.

[9]. Mithun Das, Saurabh Pandey, Shivansh Sethi, Punyajoy Saha, and Animesh Mukherjee. 2024. Low-Resource Counter speech Generation for Indic Languages: The Case of Bengali and Hindi. In Findings of the Association for Computational Linguistics: EACL 2024, pages 1601–1614, St. Julian's, Malta. Association for Computational Linguistics.

[10]. María Estrella Vallecillo Rodríguez, Maria Victoria Cantero Romero, Isabel Cabrera De Castro, Arturo Montejo Ráez, and María Teresa Martín Valdivia. 2024. CONAN-MT-SP: A Spanish Corpus for Counternarrative Using GPT Models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3677–3688, Torino, Italia. ELRA and ICCL.

[11]. Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024. On Zero-Shot Counter speech Generation by LLMs. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 12443–12454, Torino, Italia. ELRA and ICCL.

[12]. Saha, P., Singh, K., Kumar, A., Mathew, B., & Mukherjee, A. (2022). CounterGeDi: A controllable approach to generate polite, detoxified and emotional counter speech. ArXiv, abs/2205.04304.

[13]. Tekiroğlu, Serra Sinem, Helena Bonaldi, Margherita Fanton and Marco Guerini. "Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study." Findings (2022).

[14]. Wanzheng Zhu and Suma Bhat. 2021. Generate, Prune, Select: A Pipeline for Counter speech Generation against Online Hate Speech. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 134–149, Online. Association for Computational Linguistics.

[15] Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 899–914, Online. Association for Computational Linguistics.

[16]. Zubiaga, Irune, Aitor Soroa, and Rodrigo Agerri. "Ixa at refutes 2024: Leveraging language models for counter narrative generation." In IberLEF (Working Notes). CEUR Workshop Proceedings. 2024.

APPENDIX-A

PSUEDOCODE

```
import openai
# Set up OpenAI API key
openai.api_key = "your-api-key"

import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

def find_label(hate_speech_input):
  # Load dataset (change this to your dataset path)
  df = pd.read_csv('/content/master.csv')  # Replace with your file path

  # User input for hate speech
  user_input = hate_speech_input

  # Initialize the TF-IDF Vectorizer
  vectorizer = TfidfVectorizer()

  # Combine the user input with the dataset text for vectorization
  corpus = df['HS'].tolist() + [user_input]



  # Fit and transform the data to get the TF-IDF matrix
  tfidf_matrix = vectorizer.fit_transform(corpus)

  # Calculate cosine similarity between the user input (last vector) and all rows
  cosine_similarities = cosine_similarity(tfidf_matrix[-1], tfidf_matrix[:-1])

  # Add cosine similarities to the dataframe
  df['Cosine_Similarity'] = cosine_similarities.flatten()
  # Sort the dataframe by Cosine_Similarity in descending order and get top 5
  top_5 = df.nlargest(5, 'Cosine_Similarity')

  # Print the top 5 rows with the highest cosine similarity scores
  print(top_5[['Cosine_Similarity', 'HS', 'TARGET']])  # Adjust as needed for more columns

  # Check for any cosine similarity values above the threshold
  threshold = 0.7
  above_threshold = top_5[top_5['Cosine_Similarity'] > threshold]

  if not above_threshold.empty:
      # If any rows meet the threshold, the most common value is the HS value of the highest
match
      most_common_value = above_threshold['TARGET'].iloc[0]
  else:
      # Otherwise, find the most common value in the top 5 rows
      most_common_value = top_5['TARGET'].mode()[0]

  '''# Find the most common value in the 'HS' column of the top 5 rows
  most_common_value = top_5['TARGET'].mode()[0]

  # Print the most common value in the 'HS' column
  print("\nMost Common Value in the 'TARGET' Column (in Top 5):")
  print(most_common_value)'''

  if most_common_value in ['JEWS','POC','LGBT+','MIGRANTS','WOMEN']:
    return [False, most_common_value]
  else:
    return [True, most_common_value]

import openai
from langdetect import detect
from transformers import pipeline
```

```
import gradio as gr
from polyglot.detect import Detector



# Load Hugging Face Sentiment Analysis Pipeline
sentiment_analyzer = pipeline("sentiment-analysis")

def get_toxicity_score_sentiment(text):
    """
    Calculate a toxicity value based on sentiment analysis.
    Returns a toxicity value in the range [0, 1], where:
    - 0 indicates neutral/positive sentiment (low toxicity).
    - 1 indicates highly negative sentiment (high toxicity).
    """
    result = sentiment_analyzer(text)[0]  # Process the first result
    sentiment = result["label"]  # Either "NEGATIVE" or "POSITIVE"
    score = result["score"]  # Confidence score of the sentiment

    if sentiment == "NEGATIVE":
        return score  # High negative score correlates with higher toxicity
    return 1 - score  # Invert for positive/neutral sentiment

def generate_counter_narrative(hate_speech_input, mode="few-shot"):
    """
    Generate a counter-narrative for the given hate speech input.
    Supports one-shot and few-shot prompting.
    """


    # Detect the language of the input
    detector = Detector(hate_speech_input)
    detected_language = detector.language.code
    print('DETECT: ',detected_language)
    if detected_language == 'sq':
      detected_language = 'eu'
    if detected_language in ['da','sv','nl']:
      detected_language = 'en'

    language_mapping = {
        "es": "Spanish",
        "it": "Italian",
        "eu": "Basque",
        "en": "English",
    }

    if detected_language not in language_mapping:
        detected_language = 'en'

    language = language_mapping[detected_language]

    # Define one-shot and few-shot examples
    one_shot_example = f"""You are an AI trained to generate positive and informative
counter-narratives to hate speech.
    Take this as an example:
Hate speech: Transgenders don't deserve jobs.
Counter-narrative: Everyone, regardless of gender identity, has the right to work and
contribute to society. Transgender individuals possess the same skills, qualifications, and
potential as anyone else. Discriminating based on gender identity not only undermines
human dignity but also limits the rich diversity and talent that strengthens our workplaces.
Embracing inclusivity leads to stronger, more innovative, and compassionate communities.

Now, generate a counter-narrative in {language} for the following hate speech:
Hate speech: {hate_speech_input}
Counter-narrative:"""

    few_shot_example = f"""You are an AI trained to generate positive and informative
counter-narratives to hate speech.
Here are some examples:
Hate speech: Los inmigrantes están destruyendo nuestra cultura.
```

Counter-narrative: Los inmigrantes enriquecen nuestra cultura, aportando nuevas perspectivas, tradiciones y talentos que nos hacen más diversos y fuertes. La cultura de un país no es estática; se nutre y crece a través del intercambio de ideas y experiencias. Lejos de destruirla, los inmigrantes la transforman positivamente, creando una sociedad más inclusiva, abierta y vibrante.

Hate speech: Le donne non possono lavorare bene come gli uomini.
Counter-narrative: Le donne sono altrettanto capaci e competenti degli uomini in ogni ambito lavorativo. Le capacità professionali non dipendono dal genere, ma dalle esperienze, dall'educazione e dalla passione. Negare alle donne la possibilità di esprimere il loro potenziale è una perdita per la società e per il mondo del lavoro. L'uguaglianza di opportunità porta a team più forti e a risultati migliori.

Hate speech: Euskal herritarrak ez dute munduari ezer ematen.
Counter-narrative: Euskal herritarrak munduari askotariko ekarpenak egiten dizkiote, kultura, arte, zientzia eta kiroletan. Euskara, musika, dantza eta literatura, hala nola, Pablo Neruda eta Luis Iriondo bezalako pertsonaiak, mundu osoan ezagutuak dira. Euskal Herriak aniztasunaren eta berrikuntzaren sinbolo dira, eta euren ekarpenak munduko gizartean eragina izan dute.

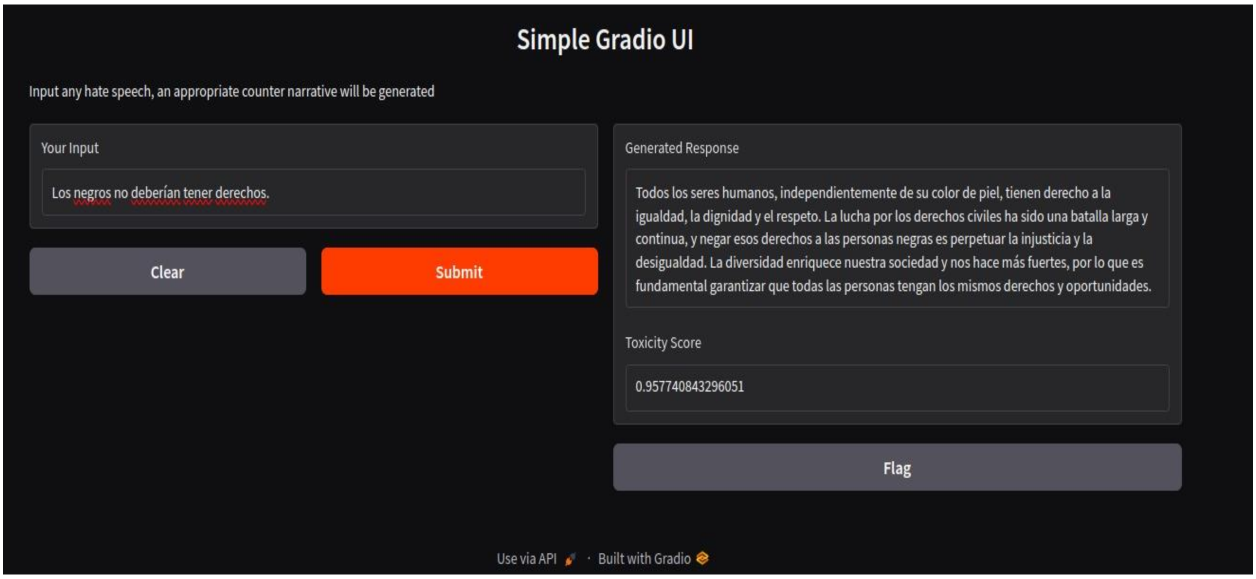
Hate speech: Immigrants are taking our jobs.
Counter-narrative: Immigrants play a vital role in the economy by filling essential roles, creating jobs, and contributing to innovation. Far from taking jobs, they often take on work that supports industries and businesses that would otherwise struggle to thrive. Immigrants also contribute to the workforce, pay taxes, and help build the communities they live in. Embracing immigration strengthens our society and boosts economic growth, benefiting everyone.

Now, generate a counter-narrative in {language} for the following hate speech:
Hate speech: {hate_speech_input}
Counter-narrative:"""

```
    # Select the prompt based on mode
    prompt = one_shot_example if mode == "one-shot" else few_shot_example
    # Generate a response using OpenAI's GPT model
    response = openai.ChatCompletion.create(
        model="gpt-4o",
        messages=[{"role": "user", "content": prompt}],
        max_tokens=128,
        temperature=0.7,
        top_p=0.95,
        frequency_penalty=0,
        presence_penalty=0.6
    )

    # Extract and return the generated text
    return response.choices[0].message['content'].strip()

def process_hate_speech_with_sentiment(hate_speech_input, mode="few-shot"):
    """
    Detect toxicity using sentiment analysis and generate a counter-narrative.
    """
    # Step 1: Calculate toxicity score based on sentiment analysis
    toxicity_score = get_toxicity_score_sentiment(hate_speech_input)

    # Step 2: Calculate toxicity score based on sentiment analysis
    default = find_label(hate_speech_input)

    # Step 3: Generate a counter-narrative
    if default[0]:
      counter_narrative = generate_default_counter_narrative(hate_speech_input,
mode=mode)
    else:
      counter_narrative = generate_counter_narrative(hate_speech_input, mode=mode)



    '''# Return both toxicity score and counter-narrative
    return counter_narrative, toxicity_score'''

    return {
        "hate_speech": hate_speech_input,
        "toxicity_score": toxicity_score,
        "label_tag": default[1],
        "counter_narrative": counter_narrative,
    }

# Create the Gradio interface
interface = gr.Interface(
    fn=process_hate_speech_with_sentiment,          # The function to process user input
    inputs=gr.Textbox(label="Your Input"),  # Input component
    outputs=[
        gr.Textbox(label="Generated Response"),
        gr.Textbox(label="Toxicity Score"),
    ],  # Output component
    title="Simple Gradio UI",       # Title for the UI
    description="Input any hate speech, an appropriate counter narrative will be generated"
)

# Launch the interface with a public URL in Colab
interface.launch(share=True)'''
```

APPENDIX-B

SCREENSHOTS



Simple Gradio UI

Input any hate speech, an appropriate counter narrative will be generated

Your Input

Los negros no deberían tener derechos.

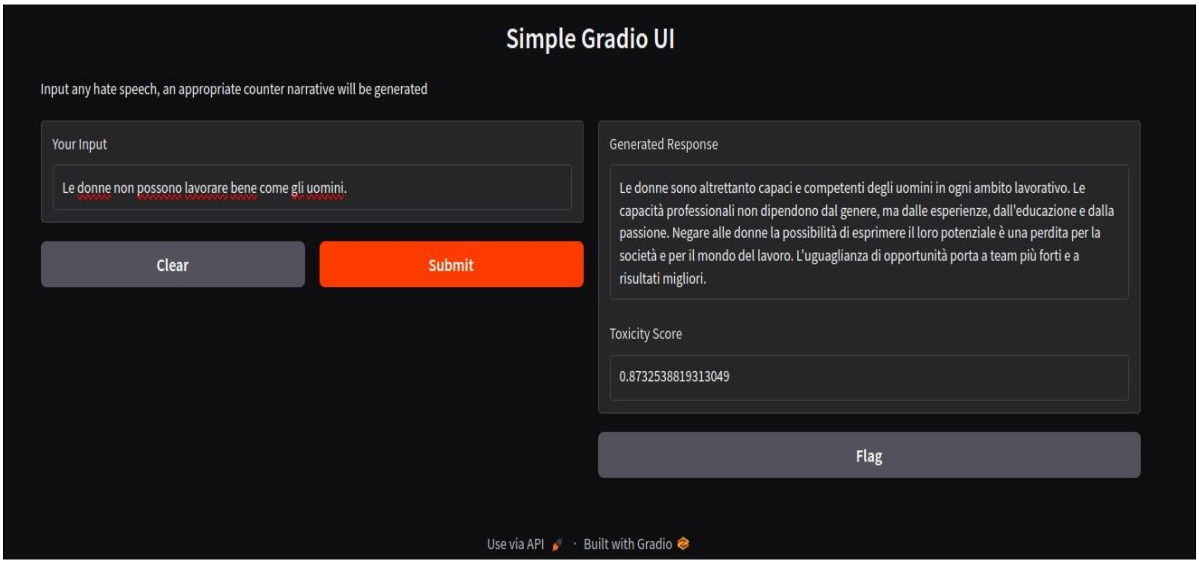
Generated Response

Todos los seres humanos, independientemente de su color de piel, tienen derecho a la igualdad, la dignidad y el respeto. La lucha por los derechos civiles ha sido una batalla larga y continua, y negar esos derechos a las personas negras es perpetuar la injusticia y la desigualdad. La diversidad enriquece nuestra sociedad y nos hace más fuertes, por lo que es fundamental garantizar que todas las personas tengan los mismos derechos y oportunidades.

Toxicity Score

0.957740843296051

Fig 13.1,
Gradio UI to input hate-speech and display a counter-narrative
for Spanish



Simple Gradio UI

Input any hate speech, an appropriate counter narrative will be generated

Your Input

Le donne non possono lavorare bene come gli uomini.

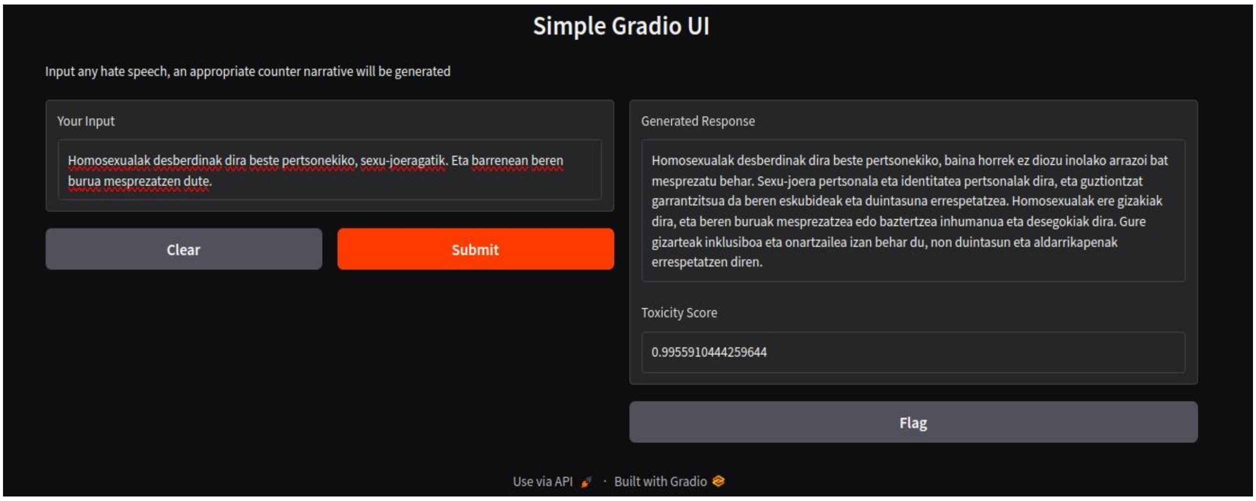
Generated Response

Le donne sono altrettanto capaci e competenti degli uomini in ogni ambito lavorativo. Le capacità professionali non dipendono dal genere, ma dalle esperienze, dall'educazione e dalla passione. Negare alle donne la possibilità di esprimere il loro potenziale è una perdita per la società e per il mondo del lavoro. L'uguaglianza di opportunità porta a team più forti e a risultati migliori.

Toxicity Score

0.8732538819313049

Fig 13.2,
Gradio UI to input hate-speech and display a counter-narrative
for Italian



Simple Gradio UI

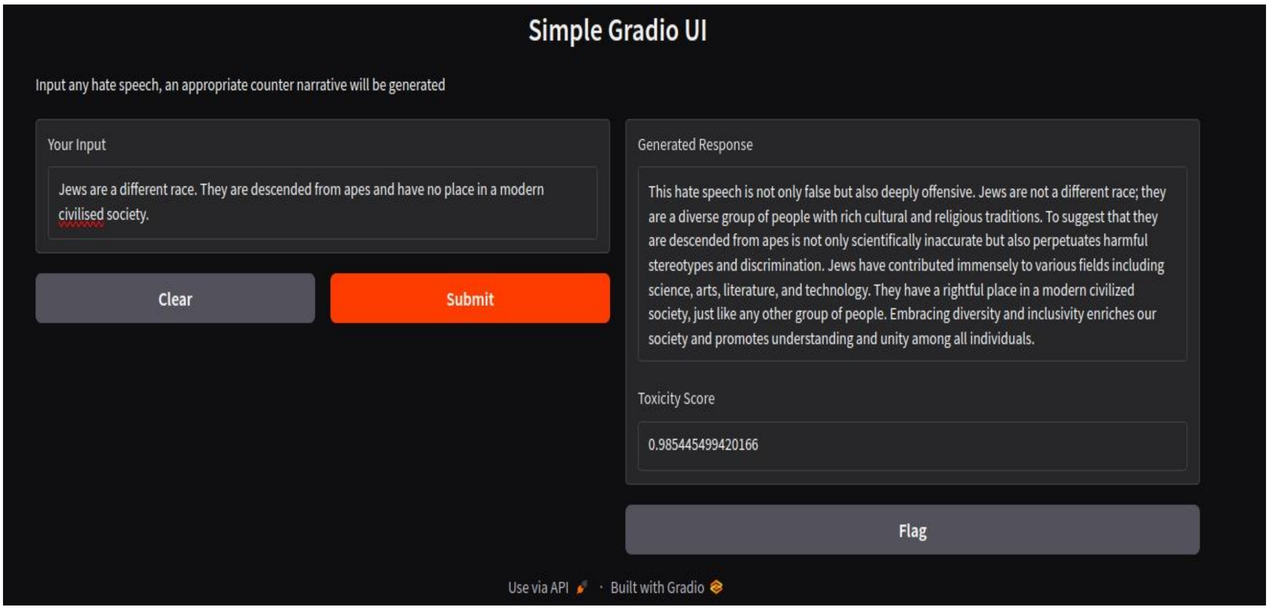
Input any hate speech, an appropriate counter narrative will be generated

Your Input: Homosexualak desberdinak dira beste pertsonekiko, sexu-joeragatik. Eta barrenean beren burua mesprezatzen dute.

Generated Response: Homosexualak desberdinak dira beste pertsonekiko, baina horrek ez diozu inolako arrazoi bat mesprezatu behar. Sexu-joera pertsonala eta identitatea pertsonalak dira, eta guztiontzat garrantzitsua da beren eskubideak eta duintasuna errespetatzea. Homosexualak ere gizakiak dira, eta beren buruak mesprezatzea edo baztertzea inhumanua eta desegokiak dira. Gure gizarteak inklusiboa eta onartzailea izan behar du, non duintasun eta aldarrikapenak errespetatzen diren.

Toxicity Score: 0.9955910444259644

Fig 13.3,
Gradio UI to input hate-speech and display a counter-narrative
for Basque



Simple Gradio UI

Input any hate speech, an appropriate counter narrative will be generated

Your Input: Jews are a different race. They are descended from apes and have no place in a modern civilised society.

Generated Response: This hate speech is not only false but also deeply offensive. Jews are not a different race; they are a diverse group of people with rich cultural and religious traditions. To suggest that they are descended from apes is not only scientifically inaccurate but also perpetuates harmful stereotypes and discrimination. Jews have contributed immensely to various fields including science, arts, literature, and technology. They have a rightful place in a modern civilized society, just like any other group of people. Embracing diversity and inclusivity enriches our society and promotes understanding and unity among all individuals.

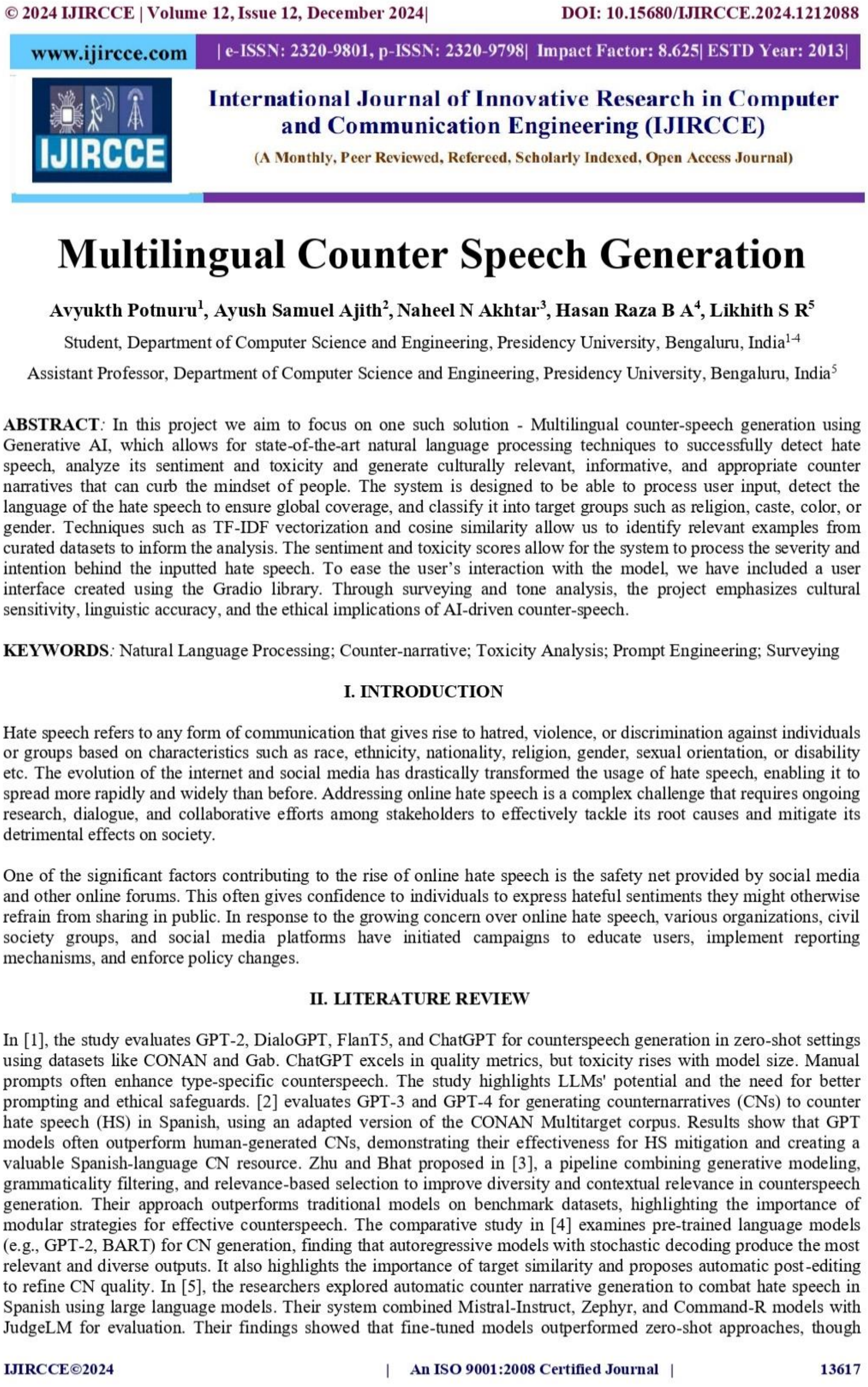
Toxicity Score: 0.985445499420166

Fig 13.4,
Gradio UI to input hate-speech and display a counter-narrative
for English

APPENDIX-C

ENCLOSURES

1. **Journal publication/Conference Paper Presented Certificates of all students.**



© 2024 IJIRCCE | Volume 12, Issue 12, December 2024| DOI: 10.15680/IJIRCCE.2024.1212088

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Multilingual Counter Speech Generation

Avyukth Potnuru[1], Ayush Samuel Ajith[2], Naheel N Akhtar[3], Hasan Raza B A[4], Likhith S R[5]

Student, Department of Computer Science and Engineering, Presidency University, Bengaluru, India[1-4]

Assistant Professor, Department of Computer Science and Engineering, Presidency University, Bengaluru, India[5]

ABSTRACT: In this project we aim to focus on one such solution - Multilingual counter-speech generation using Generative AI, which allows for state-of-the-art natural language processing techniques to successfully detect hate speech, analyze its sentiment and toxicity and generate culturally relevant, informative, and appropriate counter narratives that can curb the mindset of people. The system is designed to be able to process user input, detect the language of the hate speech to ensure global coverage, and classify it into target groups such as religion, caste, color, or gender. Techniques such as TF-IDF vectorization and cosine similarity allow us to identify relevant examples from curated datasets to inform the analysis. The sentiment and toxicity scores allow for the system to process the severity and intention behind the inputted hate speech. To ease the user's interaction with the model, we have included a user interface created using the Gradio library. Through surveying and tone analysis, the project emphasizes cultural sensitivity, linguistic accuracy, and the ethical implications of AI-driven counter-speech.

KEYWORDS: Natural Language Processing; Counter-narrative; Toxicity Analysis; Prompt Engineering; Surveying

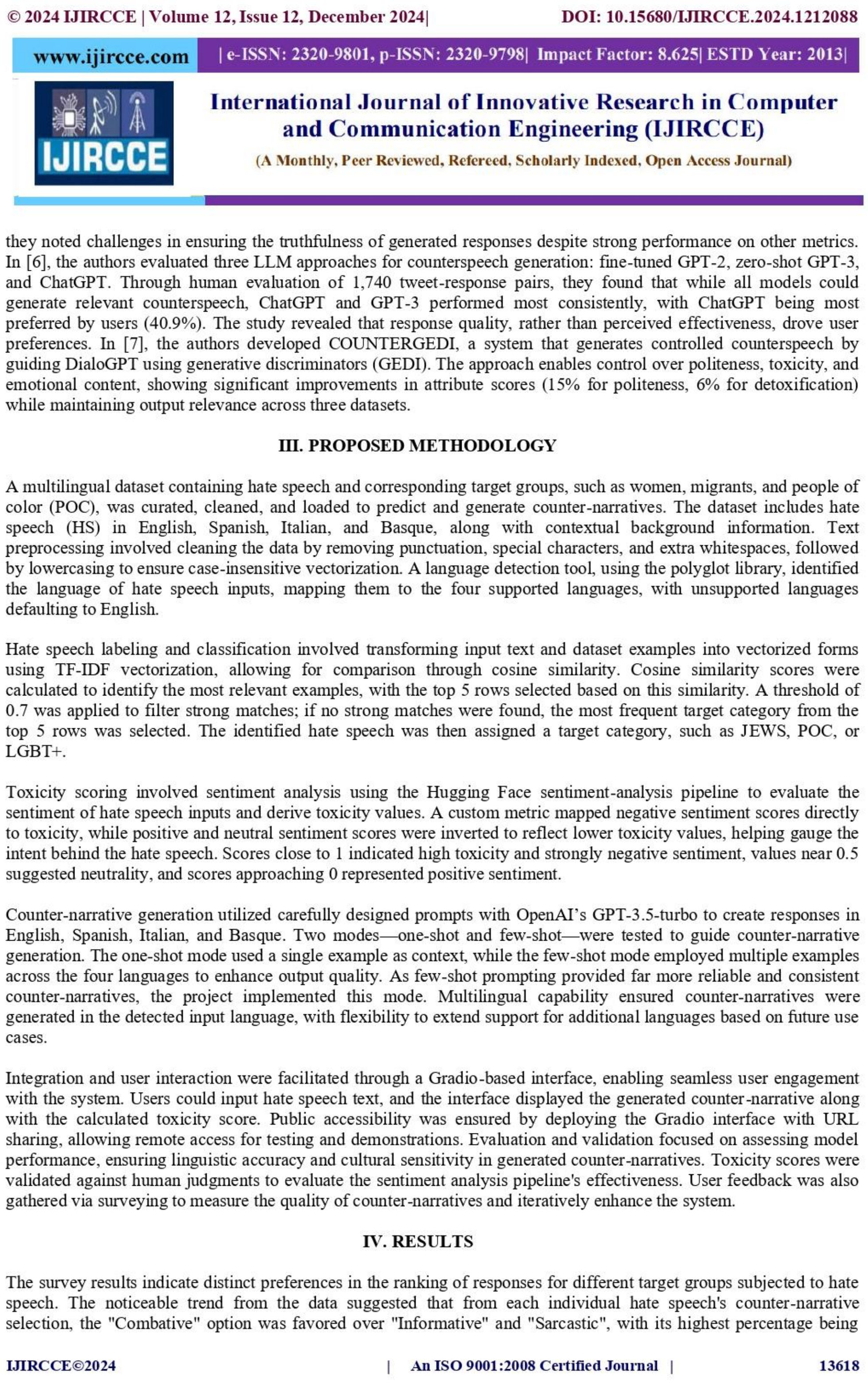
I. INTRODUCTION

Hate speech refers to any form of communication that gives rise to hatred, violence, or discrimination against individuals or groups based on characteristics such as race, ethnicity, nationality, religion, gender, sexual orientation, or disability etc. The evolution of the internet and social media has drastically transformed the usage of hate speech, enabling it to spread more rapidly and widely than before. Addressing online hate speech is a complex challenge that requires ongoing research, dialogue, and collaborative efforts among stakeholders to effectively tackle its root causes and mitigate its detrimental effects on society.

One of the significant factors contributing to the rise of online hate speech is the safety net provided by social media and other online forums. This often gives confidence to individuals to express hateful sentiments they might otherwise refrain from sharing in public. In response to the growing concern over online hate speech, various organizations, civil society groups, and social media platforms have initiated campaigns to educate users, implement reporting mechanisms, and enforce policy changes.

II. LITERATURE REVIEW

In [1], the study evaluates GPT-2, DialoGPT, FlanT5, and ChatGPT for counterspeech generation in zero-shot settings using datasets like CONAN and Gab. ChatGPT excels in quality metrics, but toxicity rises with model size. Manual prompts often enhance type-specific counterspeech. The study highlights LLMs' potential and the need for better prompting and ethical safeguards. [2] evaluates GPT-3 and GPT-4 for generating counternarratives (CNs) to counter hate speech (HS) in Spanish, using an adapted version of the CONAN Multitarget corpus. Results show that GPT models often outperform human-generated CNs, demonstrating their effectiveness for HS mitigation and creating a valuable Spanish-language CN resource. Zhu and Bhat proposed in [3], a pipeline combining generative modeling, grammaticality filtering, and relevance-based selection to improve diversity and contextual relevance in counterspeech generation. Their approach outperforms traditional models on benchmark datasets, highlighting the importance of modular strategies for effective counterspeech. The comparative study in [4] examines pre-trained language models (e.g., GPT-2, BART) for CN generation, finding that autoregressive models with stochastic decoding produce the most relevant and diverse outputs. It also highlights the importance of target similarity and proposes automatic post-editing to refine CN quality. In [5], the researchers explored automatic counter narrative generation to combat hate speech in Spanish using large language models. Their system combined Mistral-Instruct, Zephyr, and Command-R models with JudgeLM for evaluation. Their findings showed that fine-tuned models outperformed zero-shot approaches, though



© 2024 IJIRCCE | Volume 12, Issue 12, December 2024| DOI: 10.15680/IJIRCCE.2024.1212088

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

they noted challenges in ensuring the truthfulness of generated responses despite strong performance on other metrics. In [6], the authors evaluated three LLM approaches for counterspeech generation: fine-tuned GPT-2, zero-shot GPT-3, and ChatGPT. Through human evaluation of 1,740 tweet-response pairs, they found that while all models could generate relevant counterspeech, ChatGPT and GPT-3 performed most consistently, with ChatGPT being most preferred by users (40.9%). The study revealed that response quality, rather than perceived effectiveness, drove user preferences. In [7], the authors developed COUNTERGEDI, a system that generates controlled counterspeech by guiding DialoGPT using generative discriminators (GEDI). The approach enables control over politeness, toxicity, and emotional content, showing significant improvements in attribute scores (15% for politeness, 6% for detoxification) while maintaining output relevance across three datasets.

III. PROPOSED METHODOLOGY

A multilingual dataset containing hate speech and corresponding target groups, such as women, migrants, and people of color (POC), was curated, cleaned, and loaded to predict and generate counter-narratives. The dataset includes hate speech (HS) in English, Spanish, Italian, and Basque, along with contextual background information. Text preprocessing involved cleaning the data by removing punctuation, special characters, and extra whitespaces, followed by lowercasing to ensure case-insensitive vectorization. A language detection tool, using the polyglot library, identified the language of hate speech inputs, mapping them to the four supported languages, with unsupported languages defaulting to English.

Hate speech labeling and classification involved transforming input text and dataset examples into vectorized forms using TF-IDF vectorization, allowing for comparison through cosine similarity. Cosine similarity scores were calculated to identify the most relevant examples, with the top 5 rows selected based on this similarity. A threshold of 0.7 was applied to filter strong matches; if no strong matches were found, the most frequent target category from the top 5 rows was selected. The identified hate speech was then assigned a target category, such as JEWS, POC, or LGBT+.

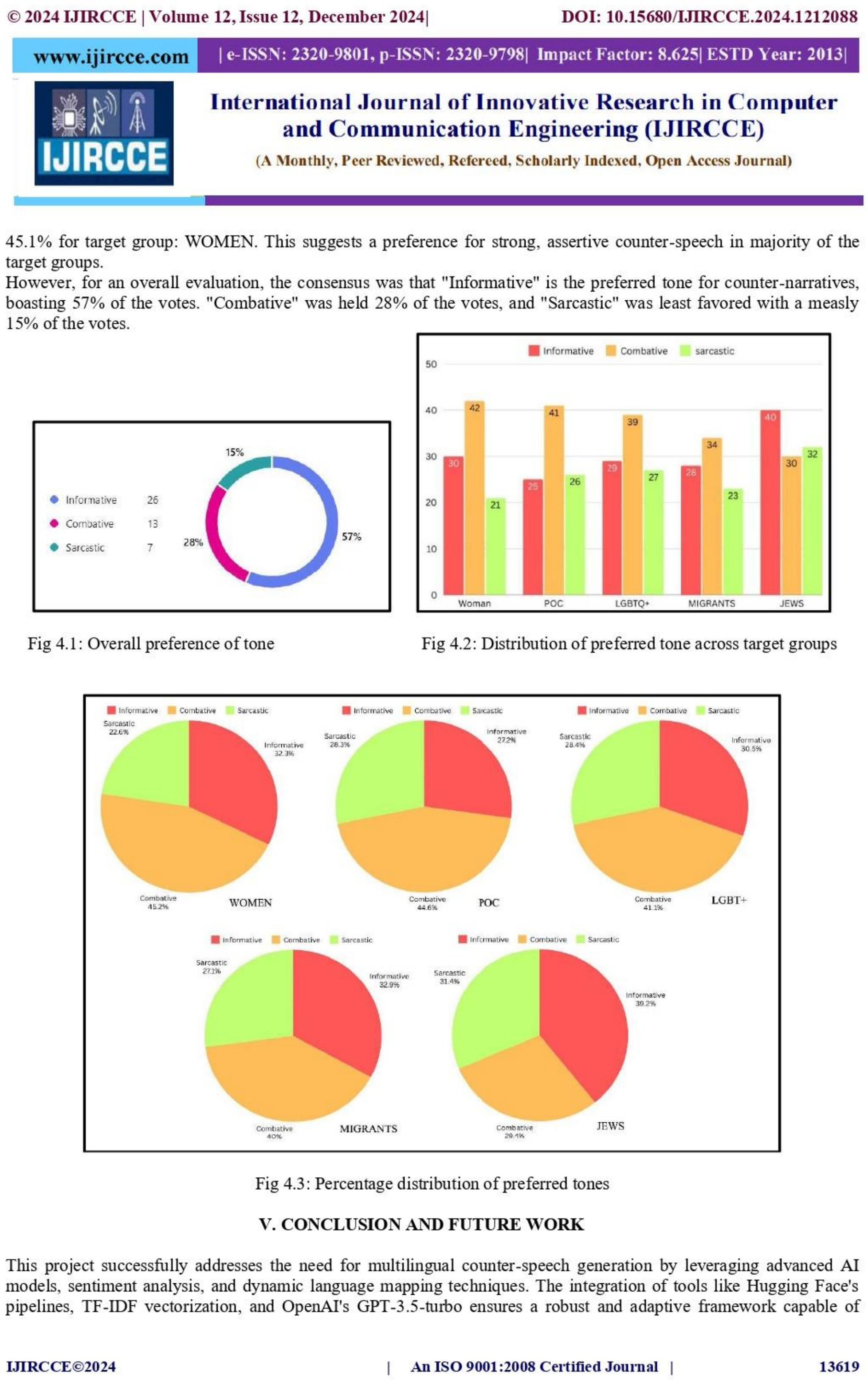
Toxicity scoring involved sentiment analysis using the Hugging Face sentiment-analysis pipeline to evaluate the sentiment of hate speech inputs and derive toxicity values. A custom metric mapped negative sentiment scores directly to toxicity, while positive and neutral sentiment scores were inverted to reflect lower toxicity values, helping gauge the intent behind the hate speech. Scores close to 1 indicated high toxicity and strongly negative sentiment, values near 0.5 suggested neutrality, and scores approaching 0 represented positive sentiment.

Counter-narrative generation utilized carefully designed prompts with OpenAI's GPT-3.5-turbo to create responses in English, Spanish, Italian, and Basque. Two modes—one-shot and few-shot—were tested to guide counter-narrative generation. The one-shot mode used a single example as context, while the few-shot mode employed multiple examples across the four languages to enhance output quality. As few-shot prompting provided far more reliable and consistent counter-narratives, the project implemented this mode. Multilingual capability ensured counter-narratives were generated in the detected input language, with flexibility to extend support for additional languages based on future use cases.

Integration and user interaction were facilitated through a Gradio-based interface, enabling seamless user engagement with the system. Users could input hate speech text, and the interface displayed the generated counter-narrative along with the calculated toxicity score. Public accessibility was ensured by deploying the Gradio interface with URL sharing, allowing remote access for testing and demonstrations. Evaluation and validation focused on assessing model performance, ensuring linguistic accuracy and cultural sensitivity in generated counter-narratives. Toxicity scores were validated against human judgments to evaluate the sentiment analysis pipeline's effectiveness. User feedback was also gathered via surveying to measure the quality of counter-narratives and iteratively enhance the system.

IV. RESULTS

The survey results indicate distinct preferences in the ranking of responses for different target groups subjected to hate speech. The noticeable trend from the data suggested that from each individual hate speech's counter-narrative selection, the "Combative" option was favored over "Informative" and "Sarcastic", with its highest percentage being



© 2024 IJIRCCE | Volume 12, Issue 12, December 2024| DOI: 10.15680/IJIRCCE.2024.1212088

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

45.1% for target group: WOMEN. This suggests a preference for strong, assertive counter-speech in majority of the target groups.

However, for an overall evaluation, the consensus was that "Informative" is the preferred tone for counter-narratives, boasting 57% of the votes. "Combative" was held 28% of the votes, and "Sarcastic" was least favored with a measly 15% of the votes.

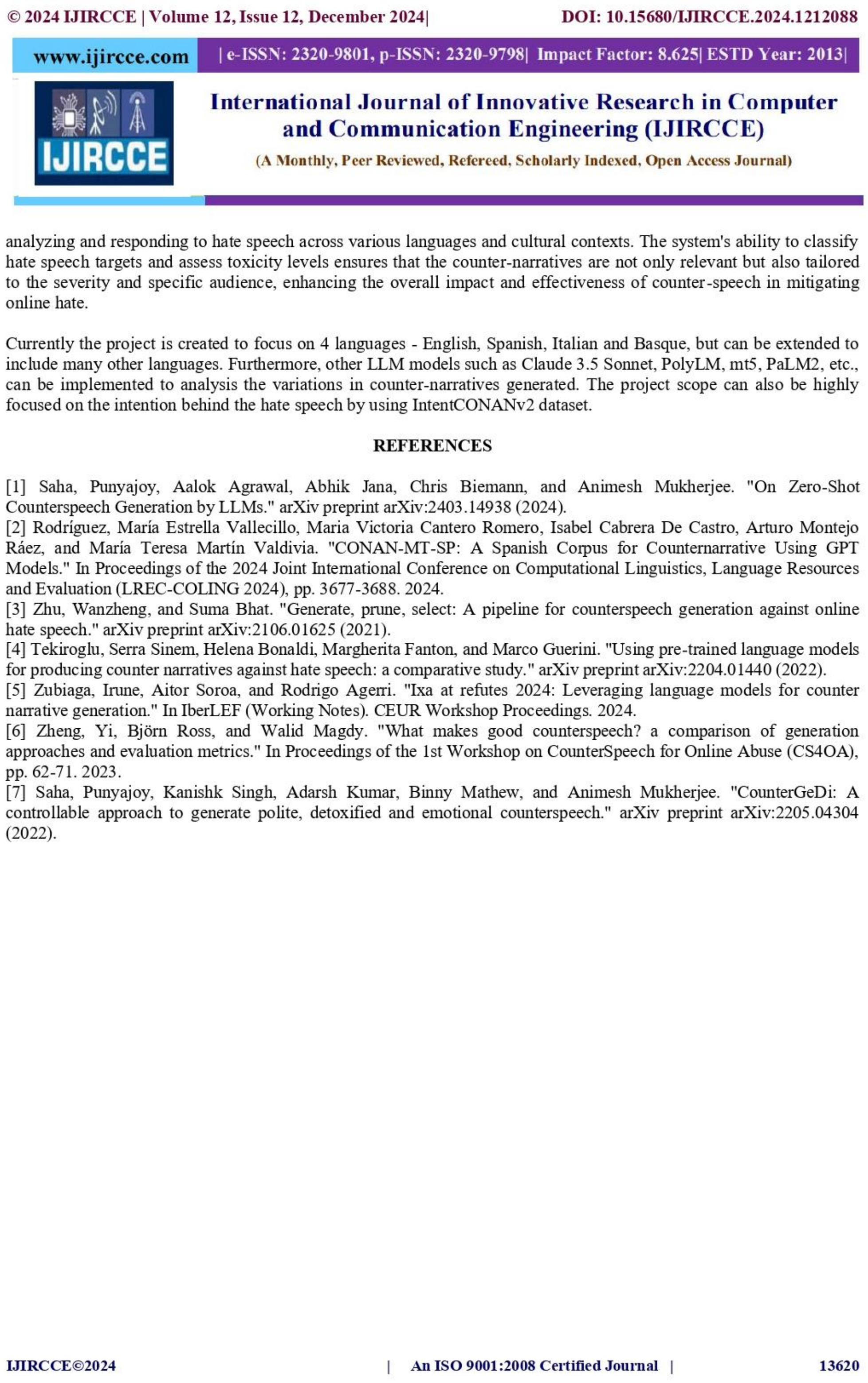
Fig 4.1: Overall preference of tone

Fig 4.2: Distribution of preferred tone across target groups

Fig 4.3: Percentage distribution of preferred tones

V. CONCLUSION AND FUTURE WORK

This project successfully addresses the need for multilingual counter-speech generation by leveraging advanced AI models, sentiment analysis, and dynamic language mapping techniques. The integration of tools like Hugging Face's pipelines, TF-IDF vectorization, and OpenAI's GPT-3.5-turbo ensures a robust and adaptive framework capable of



© 2024 IJIRCCE | Volume 12, Issue 12, December 2024| DOI: 10.15680/IJIRCCE.2024.1212088

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

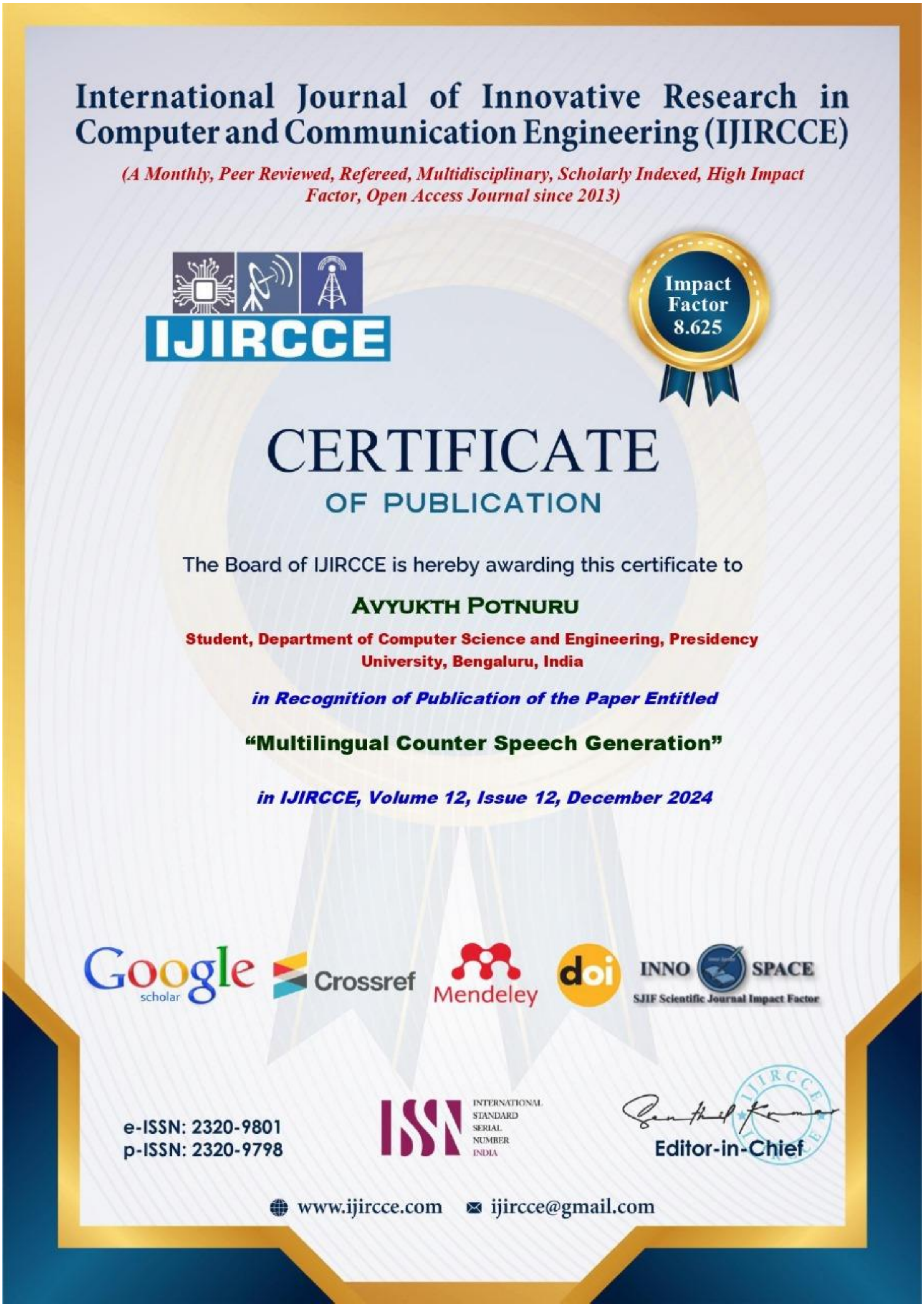
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

analyzing and responding to hate speech across various languages and cultural contexts. The system's ability to classify hate speech targets and assess toxicity levels ensures that the counter-narratives are not only relevant but also tailored to the severity and specific audience, enhancing the overall impact and effectiveness of counter-speech in mitigating online hate.

Currently the project is created to focus on 4 languages - English, Spanish, Italian and Basque, but can be extended to include many other languages. Furthermore, other LLM models such as Claude 3.5 Sonnet, PolyLM, mt5, PaLM2, etc., can be implemented to analysis the variations in counter-narratives generated. The project scope can also be highly focused on the intention behind the hate speech by using IntentCONANv2 dataset.

REFERENCES

[1] Saha, Punyajoy, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. "On Zero-Shot Counterspeech Generation by LLMs." arXiv preprint arXiv:2403.14938 (2024).
[2] Rodríguez, María Estrella Vallecillo, Maria Victoria Cantero Romero, Isabel Cabrera De Castro, Arturo Montejo Ráez, and María Teresa Martín Valdivia. "CONAN-MT-SP: A Spanish Corpus for Counternarrative Using GPT Models." In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 3677-3688. 2024.
[3] Zhu, Wanzheng, and Suma Bhat. "Generate, prune, select: A pipeline for counterspeech generation against online hate speech." arXiv preprint arXiv:2106.01625 (2021).
[4] Tekiroglu, Serra Sinem, Helena Bonaldi, Margherita Fanton, and Marco Guerini. "Using pre-trained language models for producing counter narratives against hate speech: a comparative study." arXiv preprint arXiv:2204.01440 (2022).
[5] Zubiaga, Irune, Aitor Soroa, and Rodrigo Agerri. "Ixa at refutes 2024: Leveraging language models for counter narrative generation." In IberLEF (Working Notes). CEUR Workshop Proceedings. 2024.
[6] Zheng, Yi, Björn Ross, and Walid Magdy. "What makes good counterspeech? a comparison of generation approaches and evaluation metrics." In Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), pp. 62-71. 2023.
[7] Saha, Punyajoy, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. "CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech." arXiv preprint arXiv:2205.04304 (2022).



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.625

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

AVYUKTH POTNURU

Student, Department of Computer Science and Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"Multilingual Counter Speech Generation"

in IJIRCCE, Volume 12, Issue 12, December 2024

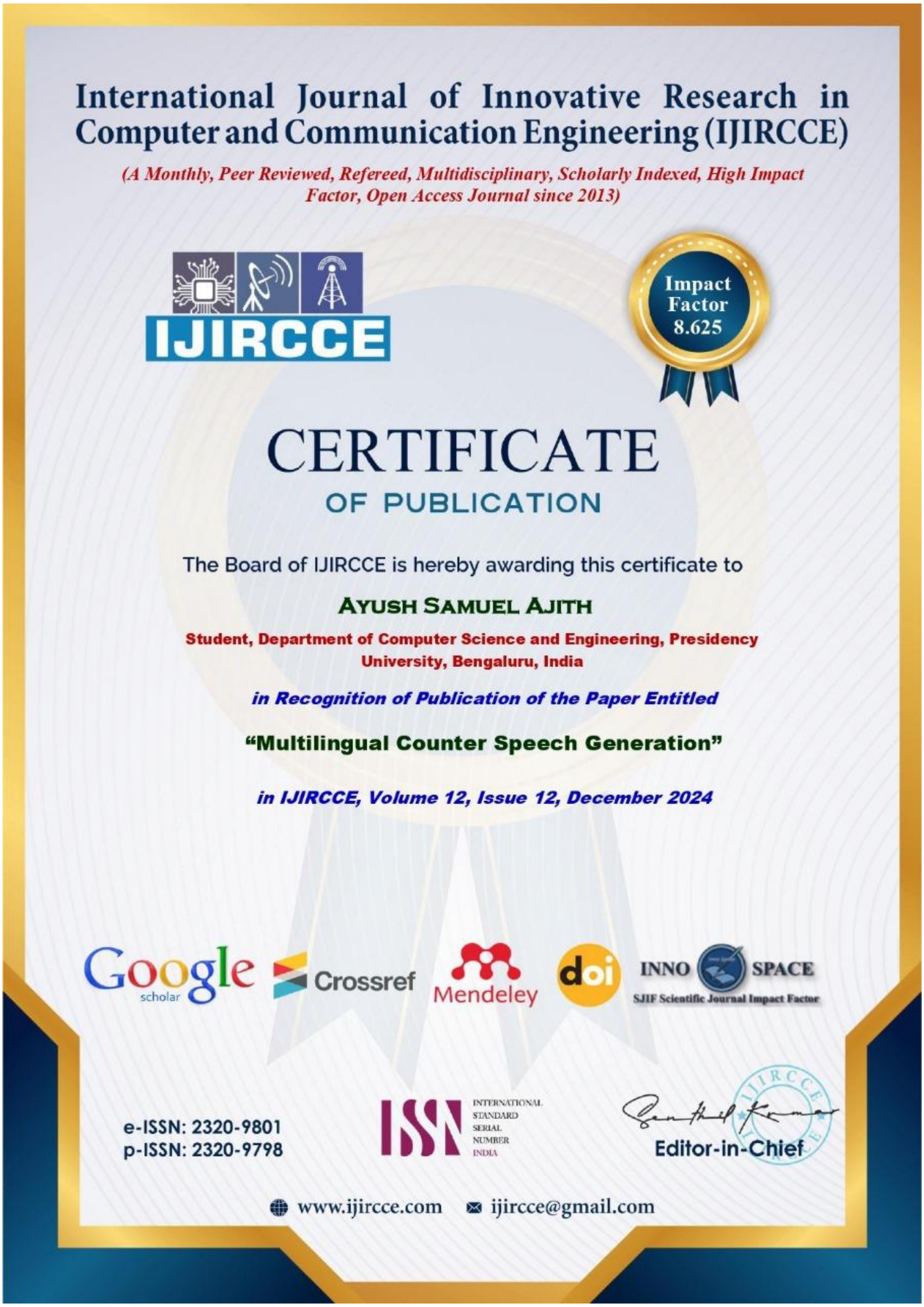
Google scholar Crossref Mendeley doi INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Senthil Kumar
IJIRCCE
Editor-in-Chief

www.ijircce.com ijircce@gmail.com



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.625

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

AYUSH SAMUEL AJITH

Student, Department of Computer Science and Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

“Multilingual Counter Speech Generation”

in IJIRCCE, Volume 12, Issue 12, December 2024

Google scholar Crossref Mendeley doi INNO SPACE SJIF Scientific Journal Impact Factor

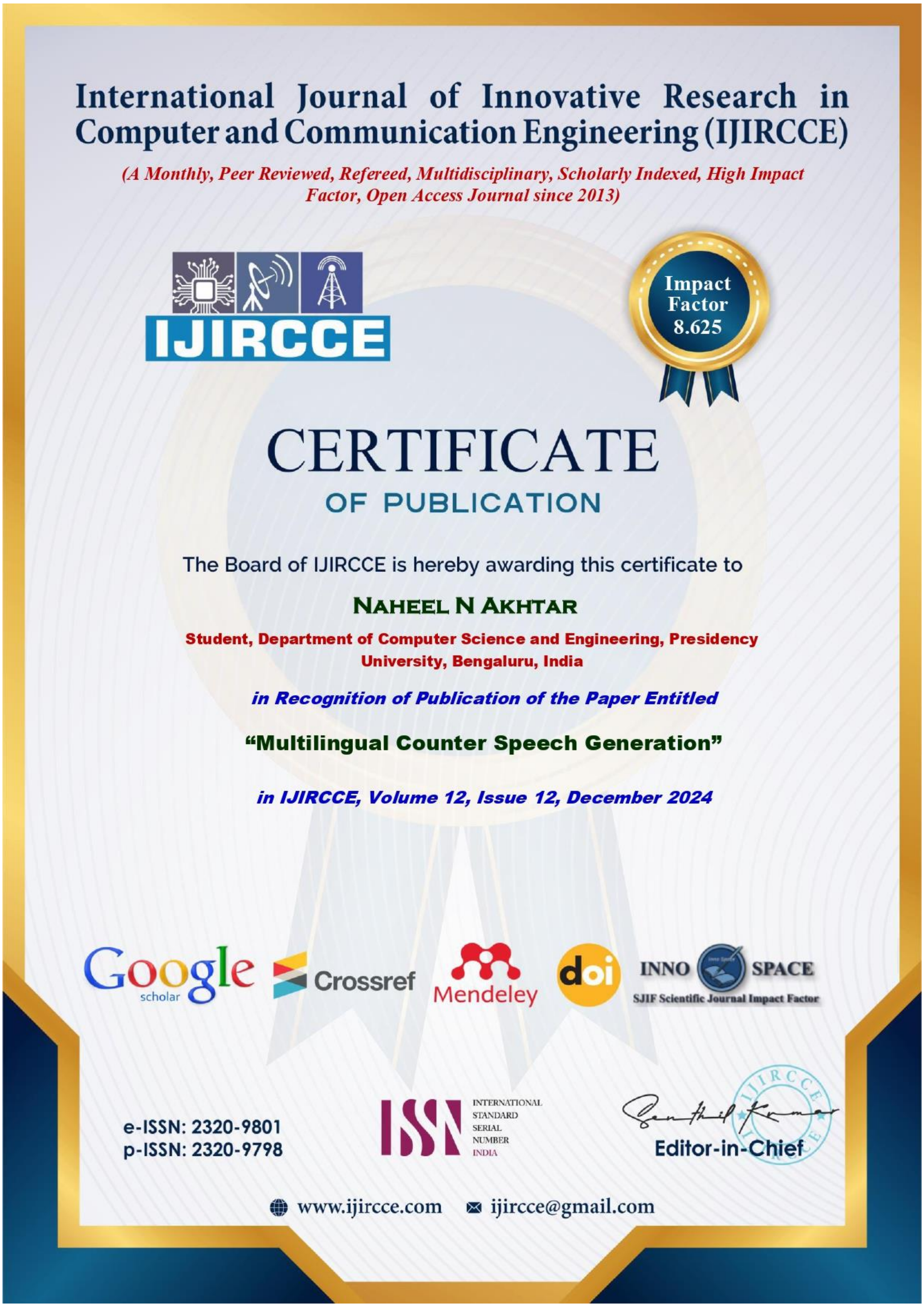
e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Senthil Kumar
Editor-in-Chief

IJIRCCE

www.ijircce.com ijircce@gmail.com



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.625

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

NAHEEL N AKHTAR

Student, Department of Computer Science and Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"Multilingual Counter Speech Generation"

in IJIRCCE, Volume 12, Issue 12, December 2024

Google scholar Crossref Mendeley doi INNO SPACE SJIF Scientific Journal Impact Factor

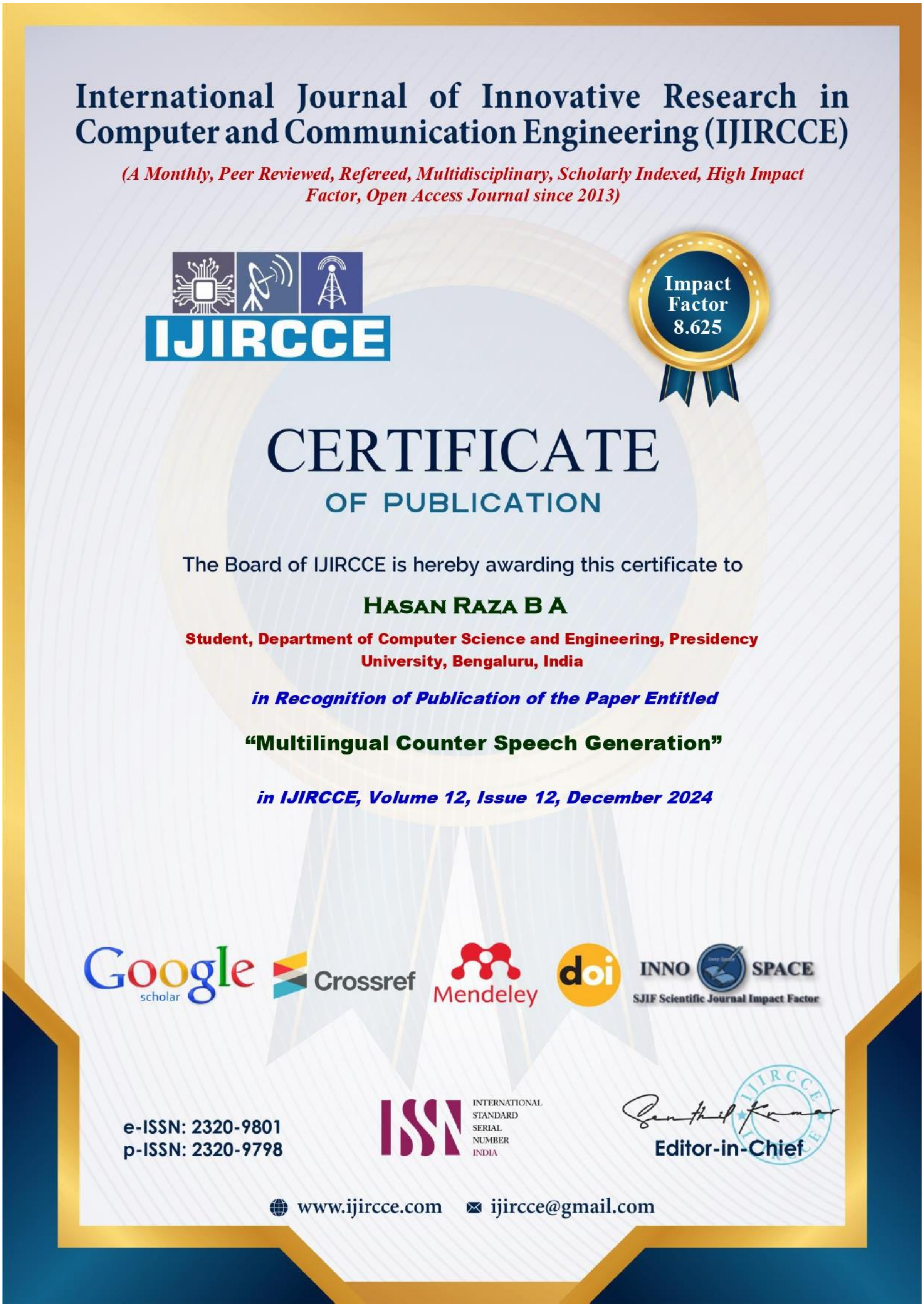
e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

IJIRCCE

Senthil Kumar
Editor-in-Chief

www.ijircce.com ✉ ijircce@gmail.com



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.625

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

HASAN RAZA B A

Student, Department of Computer Science and Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"Multilingual Counter Speech Generation"

in IJIRCCE, Volume 12, Issue 12, December 2024

Google scholar Crossref Mendeley doi INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2320-9801
p-ISSN: 2320-9798

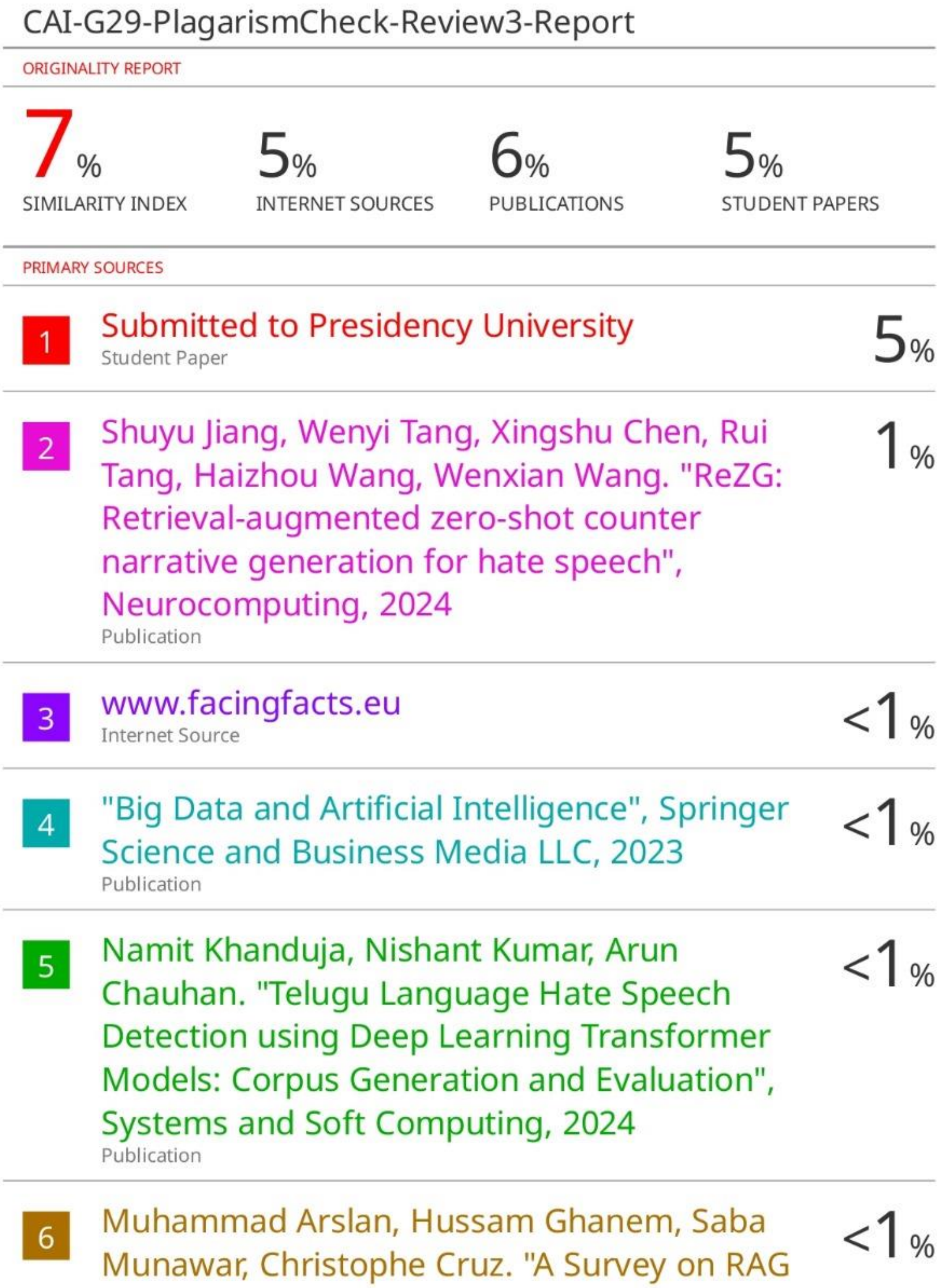
ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Senthil Kumar
Editor-in-Chief

IJIRCCE

www.ijircce.com ✉ ijircce@gmail.com

2. Similarity Index / Plagiarism Check report clearly showing the Percentage (%). No need for a page-wise explanation.



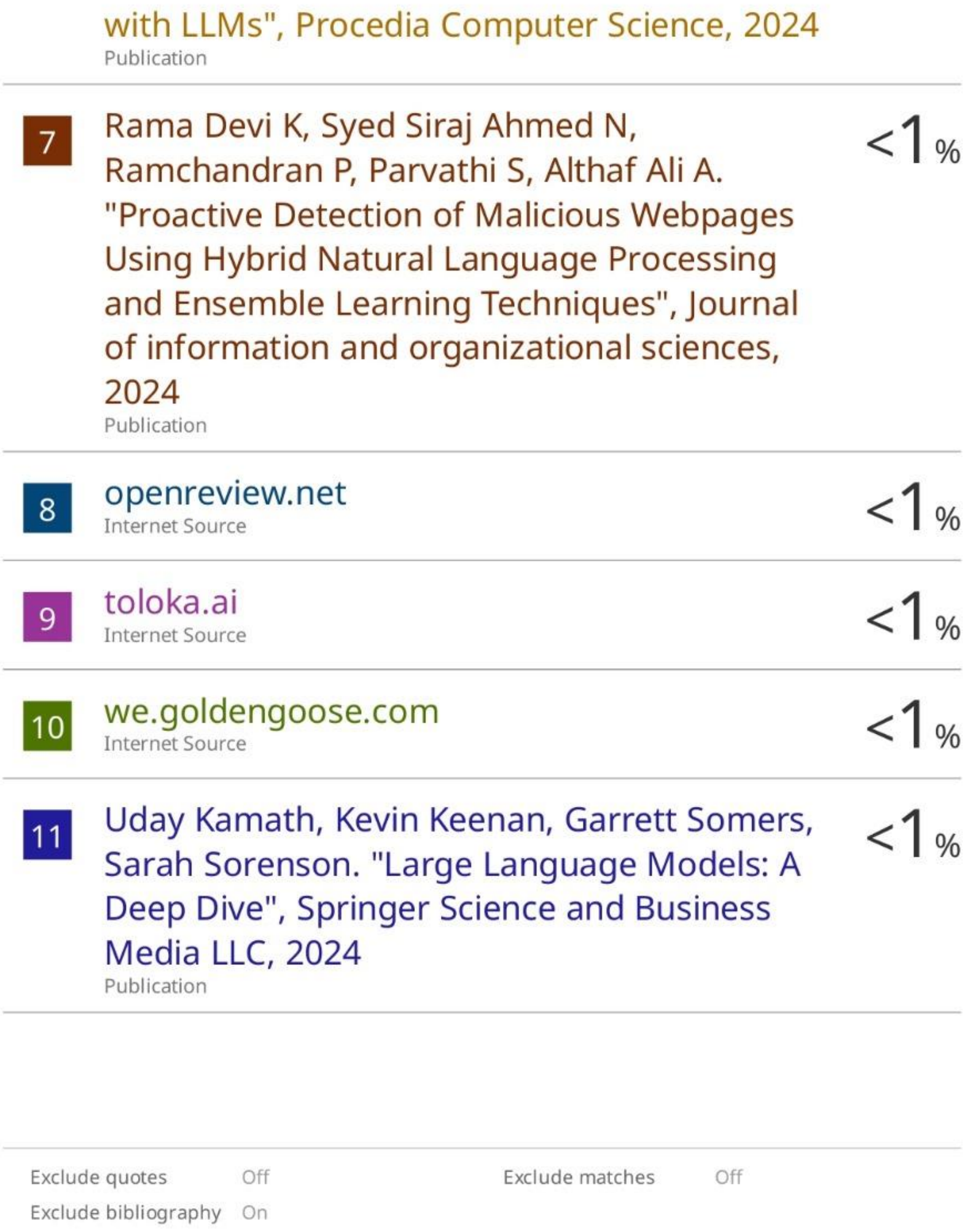
CAI-G29-PlagarismCheck-Review3-Report

ORIGINALITY REPORT

7%	5%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Presidency University Student Paper	5%
2	Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tang, Haizhou Wang, Wenxian Wang. "ReZG: Retrieval-augmented zero-shot counter narrative generation for hate speech", Neurocomputing, 2024 Publication	1%
3	www.facingfacts.eu Internet Source	<1%
4	"Big Data and Artificial Intelligence", Springer Science and Business Media LLC, 2023 Publication	<1%
5	Namit Khanduja, Nishant Kumar, Arun Chauhan. "Telugu Language Hate Speech Detection using Deep Learning Transformer Models: Corpus Generation and Evaluation", Systems and Soft Computing, 2024 Publication	<1%
6	Muhammad Arslan, Hussam Ghanem, Saba Munawar, Christophe Cruz. "A Survey on RAG	<1%



	with LLMs", Procedia Computer Science, 2024 Publication	
7	Rama Devi K, Syed Siraj Ahmed N, Ramchandran P, Parvathi S, Althaf Ali A. "Proactive Detection of Malicious Webpages Using Hybrid Natural Language Processing and Ensemble Learning Techniques", Journal of information and organizational sciences, 2024 Publication	<1 %
8	openreview.net Internet Source	<1 %
9	toloka.ai Internet Source	<1 %
10	we.goldengoose.com Internet Source	<1 %
11	Uday Kamath, Kevin Keenan, Garrett Somers, Sarah Sorenson. "Large Language Models: A Deep Dive", Springer Science and Business Media LLC, 2024 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

3. Details of mapping the project with the Sustainable Development Goals (SDGs)

SUSTAINABLE DEVELOPMENT GOALS
17 GOALS TO TRANSFORM OUR WORLD
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

- **SDG 4: Quality Education** - Counter-narratives help promote inclusive, equitable education and foster respect for diversity, contributing to peaceful discourse and critical thinking in educational contexts.
- **SDG 5: Gender Equality** - Hate speech often targets gender and sexual identity. Counter-narratives combat this by promoting gender equality and respect for women's and LGBTQ+ rights.
- **SDG 10: Reduced Inequalities** - Hate speech typically marginalizes vulnerable groups. Generating counter-narratives can address discrimination based on race, ethnicity, nationality, religion, and disability, helping to reduce social and economic inequalities.
- **SDG 16: Peace, Justice, and Strong Institutions** - Tackling hate speech fosters peaceful and inclusive societies by reducing online and offline hostility. Counter-narratives contribute to promoting justice, human rights, and combating violence, including online abuse.
- **SDG 17: Partnerships for the Goals** - Creating and disseminating counter-narratives often requires collaboration between governments, NGOs, tech platforms, and civil society, fostering partnerships that advance efforts to address hate speech at scale.