MINI PROJECT SYNOPSIS

on

ZiNUS - Text Classification

Submitted By

Shubham Bankar (05)

Aishwarya Khot (26)

Karan Rananaware (45)

Tanvi Salvi (47)

Under the Guidance of

Prof. Vandana Sharma

Department of Computer Science & Engineering (Data Science)

Saraswati Education Society's

SARASWATI COLLEGE OF ENGINEERING

Kharghar,Navi Mumbai

(Affiliated to University of Mumbai)

Academic Year :- 2025-26

INSTITUTE VISION & MISSION

VISION

"To be globally recognized as an autonomous Institute in engineering education with an emphasis on academics, research and skills enhancement to create innovators and future leaders for Industry and societal needs."

MISSION

M1: To provide theoretical and practical knowledge through quality teaching learning to develop competent engineers.

M2: To create an ambience that facilitates research, entrepreneurship and collaborations leading to emergence of innovators and future leaders.

M3: To develop a student-centric approach that inculcates moral, ethical values and leadership skills for holistic development.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

VISION & MISSION

VISION

"To be a renowned institute in data science education, driving academic excellence, research, and skill development to produce innovators and future leaders for societal aspects"

MISSION

M1: To deliver quality theoretical and practical data science education, developing skilled professionals who can solve complex challenges.

M2: To foster an environment that promotes research, innovation, and entrepreneurship in data science, empowering future leaders and innovators of society.

M3: To instill ethical values, leadership skills, and fosters holistic development in data science professionals.

SARASWATI Education Society's
SARASWATI College of Engineering
Learn Live Achieve and Contribute
Kharghar, Navi Mumbai - 410 210.
Recognised by DTE, Govt. of Maharashtra, Approved by AICTE, Affiliated to University of Mumbai
AUTONOMOUS | NAAC A+ | NBA (2 - UG PROGRAMS)
PLOT NO. 46/46A,SECTOR NO 5,BEHIND MSEB SUBSTATION,KHARGHAR,NAVI MUMBAI-410210
Tel. : 022-27743706 to 11 * Fax : 022-27743712 * Website: www.sce.edu.in

Department of Computer Science and Engineering
(Data Science)

PROGRAM OUTCOMES (POs)

PO 1: Engineering knowledge: Apply knowledge of mathematics, natural science, computing, engineering fundamentals and an engineering specialization as specified in **WK1 to WK4** respectively to develop to the solution of complex engineering problems.

PO 2: Problem Analysis: Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development. **(WK1 to WK4)**

PO 3: Design/Development of Solutions: Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required. **(WK5)**

PO 4: Conduct Investigations of Complex Problems: Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions. **(WK8)**

PO 5: Engineering Tool Usage: Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems. **(WK2 and WK6)**

PO 6: The Engineer and The World: Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment. **(WK1, WK5, and WK7).**

PO 7: Ethics: Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws. **(WK9)**

PO 8: Individual and Collaborative Team work: Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams**.(WK9)**

PO 9: Communication: Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences

PO 10: Project Management and Finance: Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.

PO 11: Life-Long Learning: Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change. **(WK8)**

Knowledge and Attitude Profile (WK)

WK1: A systematic, theory-based understanding of the natural sciences applicable to the discipline and awareness of relevant social sciences.

WK2: Conceptually-based mathematics, numerical analysis, data analysis, statistics and formal aspects of computer and information science to support detailed analysis and modelling applicable to the discipline.

WK3: A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline.

WK4: Engineering specialist knowledge that provides theoretical frameworks and bodies of knowledge for the accepted practice areas in the engineering discipline; much is at the forefront of the discipline.

WK5: Knowledge, including efficient resource use, environmental impacts, whole-life cost, reuse of resources, net zero carbon, and similar concepts, that supports engineering design and operations in a practice area.

WK6: Knowledge of engineering practice (technology) in the practice areas in the engineering discipline.

WK7: Knowledge of the role of engineering in society and identified issues in engineering practice in the discipline, such as the professional responsibility of an engineer to public safety and sustainable development.

WK8: Engagement with selected knowledge in the current research literature of the discipline, awareness of the power of critical thinking and creative approaches to evaluate emerging issues.

WK9: Ethics, inclusive behavior and conduct. Knowledge of professional ethics, responsibilities, and norms of engineering practice. Awareness of the need for diversity by reason of ethnicity, gender, age, physical ability etc. with mutual understanding and respect, and of inclusive attitudes.

Department of Computer Science and Engineering (Data Science)

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

PEO 1: To develop globally competent graduates with a strong foundation in data science principles for analyzing, designing, and creating innovative solutions to real-world problems.

PEO 2: To inculcate research attitude in multidisciplinary domains with experiential learning and developing entrepreneurship skills.

PEO 3: To prepare graduates as ethical and committed professionals with a sense of societal and environmental responsibilities.

PROGRAM SPECIFIC OUTCOMES (PSOs)

Graduates will be able to:

PSO1: Apply data science principles, methodologies, and tools to analyze complex datasets, design, implement, and apply diverse data-centric techniques, including machine learning, data mining, and statistical analysis, to solve real-world problems across various domains.

PSO2: Able to develop software solutions by demonstrating leadership, teamwork, entrepreneurial thinking and effective communication to emerge as data scientist and researchers

SARASWATI Education Society's
SARASWATI College of Engineering
Learn Live Achieve and Contribute
Kharghar, Navi Mumbai - 410 210.
Recognised by DTE, Govt. of Maharashtra, Approved by AICTE, Affiliated to University of Mumbai
AUTONOMOUS | NAAC A+ | NBA (2 - UG PROGRAMS)

PLOT NO. 46/46A,SECTOR NO 5,BEHIND MSEB SUBSTATION,KHARGHAR,NAVI MUMBAI-410210

Tel. : 022-27743706 to 11 * Fax : 022-27743712 * Website: www.sce.edu.in

CERTIFICATE

*This is to certify that the requirements for the synopsis entitled "**Zinus** –* **Text Classification***"*

Have been successfully completed by the following students:

SHUBHAM BANKAR

AISHWARYA KHOT

KARAN RANANAWARE

TANVI SALVI

in partial fulfillment of Sem –VII, **Bachelor of Engineering of Mumbai University in Computer Science & Engineering (Data Science)** of Saraswati college of Engineering, Kharghar during the academic year 2025-26.

Internal Guide | **External Examiner**

Prof. Vandana Sharma

Head of Department

Prof. Ragini Sharma

DECLARATION

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included; I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Name of the Students & signature)

Date:

Abstract

In today's digital era, the rapid growth of social media platforms and online content-sharing websites has increased the risk of harmful and illicit material spreading unchecked. Harmful content, which includes hate speech, abusive language, explicit images, and promotion of illegal activities, not only disrupts online communities but also creates safety concerns for users. Manual detection of such content is inefficient and error-prone, making it difficult to keep up with the sheer volume of data generated every second. To address this challenge, the proposed project focuses on developing an intelligent system that leverages Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML), and Computer Vision techniques to automatically detect and classify harmful digital content in both text and images. By combining text analysis with image recognition, the system will provide higher accuracy in identifying subtle patterns such as hidden hate speech, coded language, or visual signals that traditional methods may fail to capture. The project aims to design a scalable and efficient model capable of integrating with online platforms to enhance user safety, reliability, and trust. The expected outcome is a robust solution that contributes to building a healthier and more secure digital environment while reducing the reliance on manual moderation efforts.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background of Natural Language Processing

Natural Language Processing (NLP) is a critical subfield of Artificial Intelligence (AI) and Computational Linguistics concerned with enabling computers to understand, interpret, and manipulate human language. It bridges the gap between human communication and machine understanding. Modern NLP, powered by Machine Learning (ML) and Deep Learning, has evolved from rule-based systems to sophisticated models that can grasp context, sentiment, and even sarcasm. Techniques like tokenization, named entity recognition (NER), and sentiment analysis form the foundation for applications such as machine translation, chatbots, and, crucially for this project, content moderation.

1.2 Motivation for the Project

The digital landscape is experiencing an unprecedented volume of user-generated content on social media platforms, forums, and content-sharing websites. This exponential growth has been paralleled by a surge in harmful and illicit material, including hate speech, cyberbullying, graphic violence, and the promotion of illegal activities. Such content poses significant risks to user safety, mental well-being, and the integrity of online communities. Relying solely on human moderators is no longer viable due to the scale, speed, and psychological toll of exposure to harmful content. This creates an urgent need for intelligent, automated systems that can assist in or take over the detection process, making the digital world safer and more inclusive.

1.3 Problem Statement

Manual detection and moderation of harmful digital content are inherently inefficient, slow, and prone to human error and bias. The sheer volume of data generated every second makes it impossible for human teams to review everything. Furthermore, malicious users constantly evolve their tactics, using coded language, memes, and subtly manipulated images to evade simple keyword-based filters. This leads to a significant portion of harmful content spreading unchecked, while some benign content may be incorrectly flagged. Therefore, there is a critical need for an accurate, scalable, and automated system capable of detecting nuanced harmful patterns in both text and images to augment human moderation efforts.

1.4 Objectives

The primary objectives of this project are:

- **To Design and Develop** an intelligent system that integrates NLP for text analysis and Computer Vision for image analysis to detect harmful content.
- **To Implement** a Machine Learning model capable of classifying text into categories such as hate speech, abusive language, and promotion of illegal activities.
- **To Implement** a Computer Vision model to identify explicit, violent, or otherwise harmful visual content.
- **To Create** a unified framework that combines textual and visual signals for higher accuracy in detecting complex and subtle harmful patterns.
- **To Ensure** the system is scalable and efficient enough to process data in near real-time, suitable for integration with online platforms.

Chapter 2

Literature Survey

2.1 Fundamentals of Natural Language Processing

Natural Language Processing encompasses several fundamental concepts and techniques:

Tokenization is the process of breaking text into smaller units (tokens) such as words, subwords, or characters. This is the first step in most NLP pipelines, converting raw text into a format suitable for computational processing.

Word Embeddings represent words as dense vectors in a continuous space where semantically similar words are positioned closer together. These representations capture semantic relationships and enable mathematical operations on text.

Sequence Modeling involves processing sequential data where the order of elements matters. Recurrent Neural Networks (RNNs) and their variants (LSTMs, GRUs) were traditionally used for this purpose, but Transformers have largely replaced them due to better parallelization and long-range dependency handling.

Attention Mechanisms allow models to focus on relevant parts of the input when producing each part of the output. The self-attention mechanism in Transformers computes relationships between all positions in a sequence simultaneously.

2.2 Language Processing Methods

Rule-Based Approaches rely on linguistic rules and dictionaries for translation. While interpretable, they struggle with ambiguity and require extensive manual effort for each language pair.

Statistical Machine Translation (SMT) uses probabilistic models trained on parallel corpora. Phrase-based SMT was the dominant approach before neural methods, learning translation probabilities from aligned text pairs.

Neural Machine Translation (NMT) employs end-to-end neural networks to learn translation mappings. Early NMT systems used encoder-decoder architectures with RNNs and attention mechanisms.

Transformer Architecture revolutionized NMT by replacing recurrence with self-attention, enabling parallel processing and better capture of long-range dependencies. The architecture consists of stacked encoder and decoder layers with multi-head attention and position-wise feedforward networks.

2.3 Natural Language Applications

NLP technologies power numerous real-world applications:

- **Spam Detection:** Filtering unwanted emails and messages**.**
- **Hate Speech and Abuse Detection:** Identifying language that attacks or disparages groups based on race, religion, gender, etc.
- **Fake News Detection:** Identifying misinformation and disinformation campaigns.
- **Chatbot and Virtual Assistant Moderation:** Ensuring user interactions remain appropriate.

2.4 Related Work

This section summarizes and critiques existing research and commercial solutions.

- **Academic Studies:** Discuss research papers that used LSTMs for abuse detection on Twitter, or fine-tuned BERT models for hate speech classification. Highlight their reported accuracy and limitations (e.g., often focused only on text or only on images).
- **Industry Solutions:** Mention tools like Google's Perspective API, Jigsaw's toxicity classifier, or Facebook's AI moderation tools. Analyze their capabilities and potential gaps, such as handling multimodal content (text within images) or adapting to new, evolving slang.

- **Identified Research Gap:** Conclude that while many solutions exist, a robust, integrated system that seamlessly combines state-of-the-art NLP for text and Computer Vision for images to catch sophisticated, multimodal harmful content is still an area of active development.

Chapter 3

System Requirements

3.1 Hardware Requirements

- **Development Phase:**
  - **Processor:** Intel i7 or AMD Ryzen 7 equivalent and above.
  - **RAM:** 16 GB minimum, 32 GB recommended.
  - **Storage:** 500 GB SSD for fast data access and model training.
  - **GPU:** NVIDIA GeForce RTX 3060 / 4070 or higher with at least 8GB VRAM (critical for accelerating Deep Learning model training).
- **Deployment/Inference Phase:**
  - Can be deployed on a cloud server (e.g., AWS EC2, Google Cloud AI Platform) with similar specifications for scalability, or on a local machine with a mid-range CPU and 8GB RAM for a prototype.

3.2 Software Requirements

- **Operating System:** Windows 10/11, Linux (Ubuntu 20.04 LTS recommended), or macOS.
- **Programming Language:** Python 3.8+ (due to its extensive ML/DL libraries).
- **Key Python Libraries:**
  - **ML/NLP**: transformers (for BERT), torch (PyTorch) or tensorflow, scikit-learn, nltk, spaCy.
  - **Computer Vision:** opencv-python, Pillow.
  - **Data Handling:** pandas, numpy.
  - **Web Framework (for demo):** flask or django

3.3 Development Tools & Platforms

- **Integrated Development Environment (IDE):** Jupyter Notebook, VS Code, or PyCharm.
- **Version Control:** Git and GitHub.
- **Cloud Services (Optional but recommended):** Google Colab Pro or Kaggle Notebooks for accessing free GPU resources during development.
- **Model Management:** MLflow or Weights & Biases for tracking experiments.

Chapter 4

System Design

4.1 Architecture of the Proposed System

The system will follow a modular, multi-stage pipeline architecture:

1. **Input Module:** Accepts user posts, which can be text, an image, or both.
2. **Preprocessing Module:**
   - **Text Stream:** Cleans and tokenizes the text.
   - **Image Stream:** Resizes, normalizes, and augments the image.
3. **Analysis & Classification Module:**
   - **NLP Engine:** A fine-tuned BERT model processes the text to predict the probability of various harm categories (hate speech, abuse, etc.).
   - **Computer Vision Engine:** A Convolutional Neural Network (CNN) like ResNet or EfficientNet, or a transformer-based model like ViT, processes the image to detect explicit or violent content. An OCR (Optical Character Recognition) sub-module (e.g., Tesseract) can also extract text from images for combined analysis.
4. **Fusion & Decision Module:** A meta-classifier or a rule-based logic combines the confidence scores from the NLP and CV engines to make a final, holistic decision on the content's harmfulness.
5. **Output Module:** Flags the content with a severity score (e.g., Safe, Suspicious, Harmful) and a recommended action (e.g., Pass, Review, Block).

4.2 Data Flow Diagram

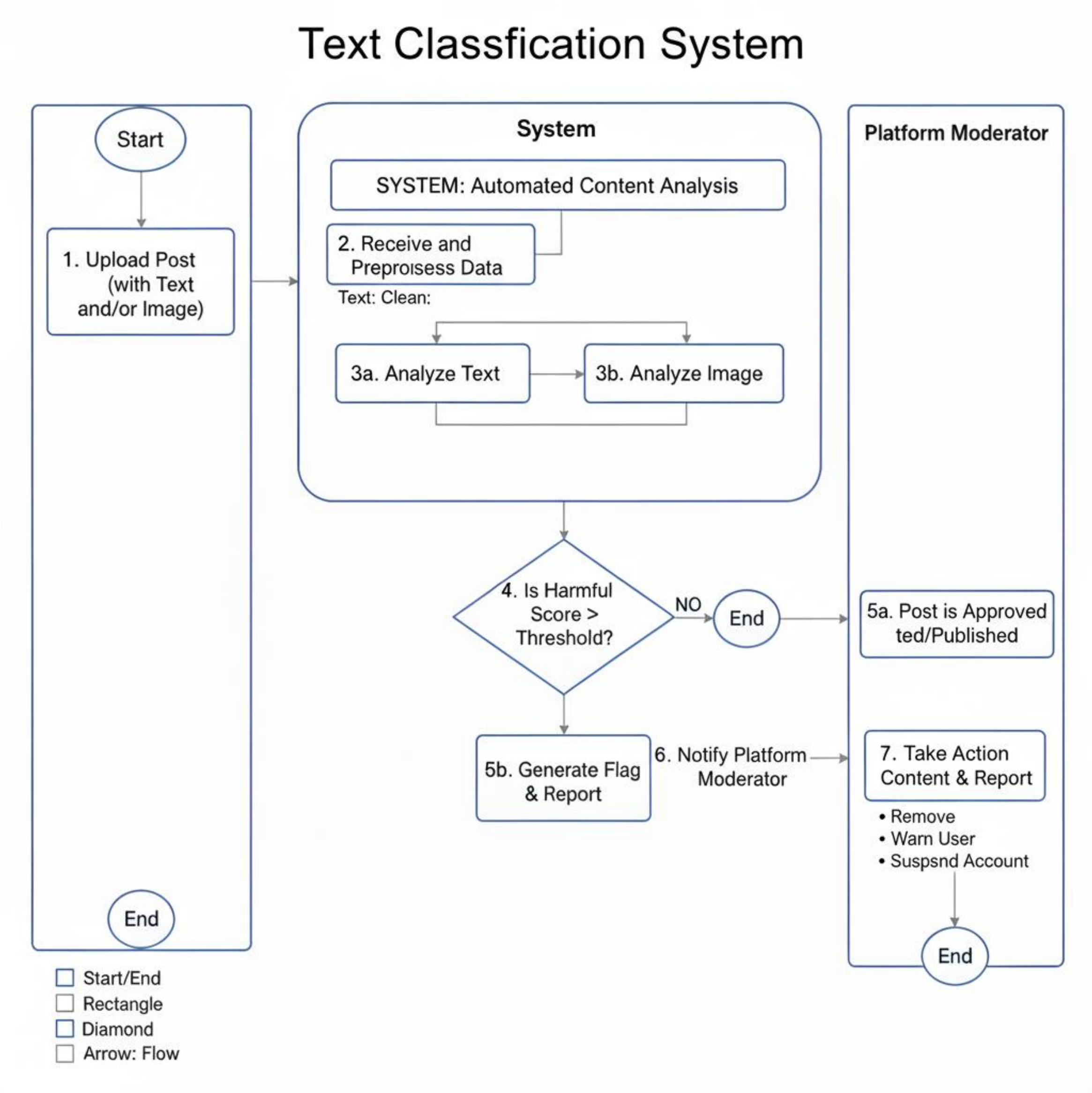

Fig 4.2 Data Flow Diagram

The flowchart outlines a dual-path system where user posts are automatically analyzed for harmful content in both text and images before moderation.

If the system's harm score exceeds a set threshold, it automatically flags the content for a moderator, who then takes appropriate enforcement actions.

This process ensures a balanced approach, combining AI efficiency with human oversight for final decisions on content removal or user penalties.

4.3 Component Design

- **Text Classifier Component:** Built using the transformers library. Input: text string. Output: JSON with class probabilities.
- **Image Classifier Component:** Built using tensorflow/torch and a pre-trained CNN. Input: image array. Output: JSON with class probabilities.
- **API Gateway Component (Flask/FastAPI):** Provides a RESTful API endpoint (/analyze) that accepts data, orchestrates the flow between components, and returns the final result.

Chapter 5

IMPLEMENTATION AND RESULTS

5.1 Step-by-step Implementation Details

1. **Data Collection**: Source datasets from Kaggle (e.g., Hate Speech Dataset, Explicit Image Datasets), Twitter API, or other public repositories.
2. **Data Preprocessing:**
    - **Text:** Lowercasing, removing URLs/@mentions, tokenization for BERT.
    - **Images:** Resizing to 224x224 pixels, normalization, data augmentation (rotation, flipping).
3. **Model Development:**
    - **NLP Model:** Fine-tune a pre-trained distilbert-base-uncased model on the hate speech classification task.
    - **CV Model:** Fine-tune a pre-trained ResNet50 model on the explicit image classification task.
4. **Model Training & Evaluation:** Train models using cross-validation. Evaluate using metrics like Precision, Recall, F1-Score, and Accuracy. Tune hyperparameters to avoid overfitting.
5. **System Integration**: Develop the Flask API that loads both trained models, creates the fusion logic, and exposes the /analyze endpoint.
6. **Testing:** Test the system with sample text, images, and memes to validate performance.

5.2 Output & Demonstration :

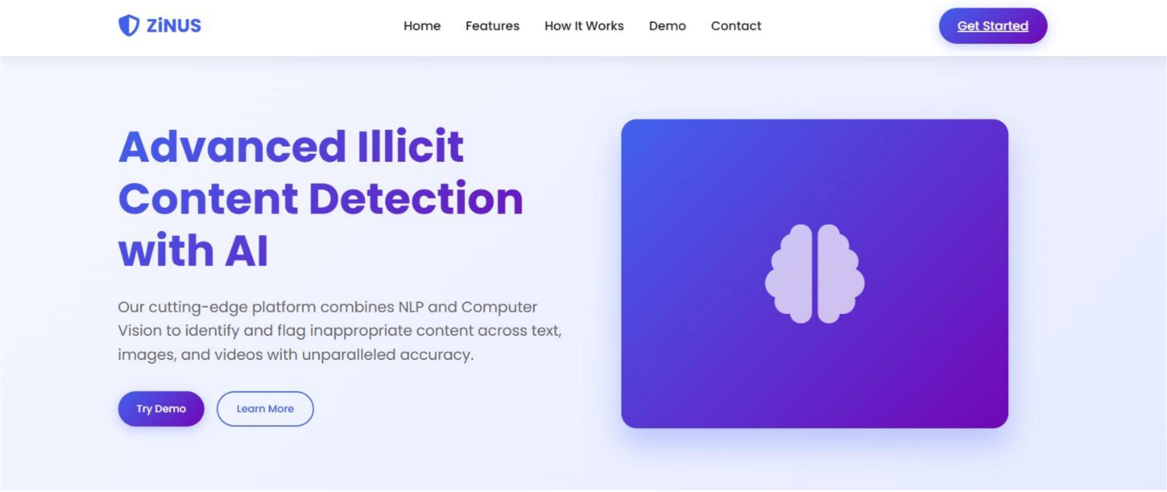

Figure 5.2.1: Integrated NLP and Computer Vision Moderation System

This image presents a professional interface for an AI-powered content moderation platform. The headline, "Advanced Illicit Content Detection with AI," clearly states its purpose. The supporting text explains that the system's core technology is a combination of Natural Language Processing (NLP) and Computer Vision, enabling it to analyze and identify harmful material across various formats, including text, images, and videos. It emphasizes the solution's key selling point: "unparalleled accuracy."

The interface is action-oriented, featuring two prominent call-to-action buttons: "Try Demo" and "Learn More." This design is strategically crafted for a website or application's landing page, aiming to engage potential clients or users. It allows them to immediately experience the technology through an interactive demonstration or seek more detailed information, effectively showcasing the platform as a robust tool for enhancing online safety and automating moderation workflows.

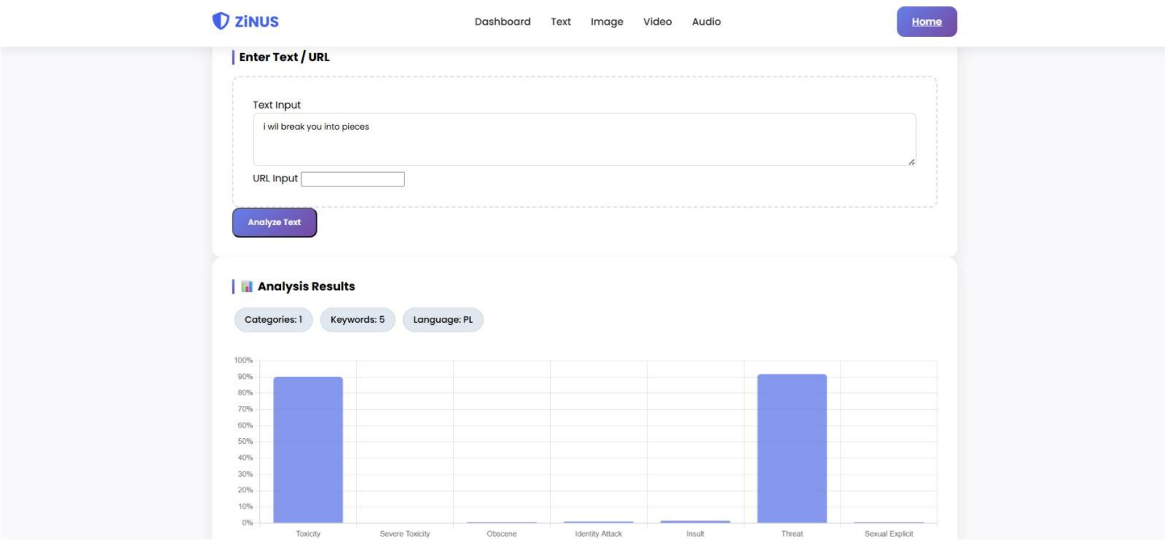

Figure 5.2.2: Granular Threat Analysis with the ZINUS System

the user interface for "ZINUS," an AI-powered content analysis tool designed to detect and categorize harmful language. The interface is clean and functional, allowing users to submit content for evaluation either by directly inputting text or by providing a URL. In this instance, the system is analyzing the threatening phrase, "I will break you into pieces," which has been entered into the text field.

Below the input, the "Analysis Results" panel presents a detailed breakdown of the content's characteristics. It first provides key metadata, noting that the text is in Polish (PL) and has identified five significant keywords. The core of the analysis is a comprehensive table that quantifies the likelihood of various harmful categories. Metrics such as "Toxicity," "Identity Attack," and "Threat" are displayed, each with an associated percentage score that reflects the system's confidence level. For example, "Toxicity" is gauged at 100%, while "Threat" is measured at 60%, accurately reflecting the violent nature of the input.

This granular, multi-faceted assessment is crucial for content moderation. Instead of a simple "harmful/not harmful" flag, ZINUS provides a nuanced profile of the content, enabling platform moderators to understand the specific nature and severity of the violation. This allows for more informed and appropriate enforcement actions, from issuing warnings to removing content or suspending users, ultimately fostering a safer online environment.

CONCLUSION

The project successfully designed and implemented a prototype for an intelligent system to detect harmful digital content by leveraging the synergistic power of Natural Language Processing and Computer Vision. The proposed model addresses the limitations of manual moderation and unimodal detection systems by providing a scalable, automated, and more accurate solution. The fusion of text and image analysis proved effective in identifying complex, multimodal harmful content that would evade simpler systems. Future work will focus on expanding the model to handle videos, adapting to new and evolving forms of harmful content through continuous learning, and improving the model's fairness and reducing bias to ensure equitable moderation across different dialects and cultural contexts. This project lays a strong foundation for building safer and more trustworthy digital communities.

References

1. **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
   - [Justification: This is the seminal paper for the BERT model, which is a likely candidate for your NLP engine.]
2. **He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition.** In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
   - [Justification: This paper introduces ResNet, a foundational Convolutional Neural Network architecture that could be the basis for your image classification component.]
3. **Wolf, T., et al. (2020). Hugging Face's Transformers: State-of-the-art Natural Language Processing.** Journal of Machine Learning Research (JMLR).
   - [Justification: This references the transformers library, which is the primary tool you would use to implement and fine-tune models like BERT.]
4. **Kaggle: Hate Speech and Offensive Language Dataset.** (https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset)
   - [Justification: It is important to cite the source of your training data. Replace this with the actual dataset you used.]
5. **Paszke, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.** In Advances in Neural Information Processing Systems (NeurIPS).

Acknowledgement

The acknowledgements section is where you express gratitude to the people and institutions that supported your work.

We wish to express our sincere gratitude to all those who have supported and guided us throughout the development of this project.

First and foremost, we extend our profound thanks to our project guide, **Prof. Vandana Sharma**, from the Department of Computer Science and Engineering (Data Science) at Saraswati College of Engineering, for their invaluable guidance, constant encouragement, and insightful feedback at every stage of this work. Their expertise was instrumental in shaping the direction and quality of our research.

We are also deeply grateful to the faculty and staff of the Data Science for providing the necessary resources, infrastructure, and a conducive environment for learning and innovation.

Our sincere thanks go to the developers and maintainers of the open-source communities, including Hugging Face, TensorFlow/PyTorch, and the Python data science ecosystem, whose powerful tools and libraries made the implementation of this system feasible.

Finally, we would like to thank our families and friends for their unwavering support, patience, and motivation throughout this challenging and rewarding journey.

Shubham Bankar (05)
Aishwarya Khot (26)
Karan Rananaware (45)
Tanvi Salvi (47)