UNICC | Artificial Intelligence

NYU | SPS

Responsible AI Innovation for Social Impact:

A Case Study in Multilingual Media Moderation

September 2025

Authors

Anusha Dandapani
Chief, AI Hub,
United Nations International Computing Centre (UNICC).

Shambhavi Mohan
Lead Machine Learning Scientist
United Nations International Computing Centre (UNICC).

Devyani Rastogi
Data Analyst Intern,
United Nations International Computing Centre (UNICC).

Dr. Andres Fortino
Clinical Associate Professor of Management and Analytics
School of Professional Studies, New York University (NYUSPS)

Siranush Kostanyan
Technical Program Manager, Building Wings
M.S. in Project Management, New York University (NYUSPS)

Editors

Jean Pierre Mora Casasola
Communications Officer
United Nations International Computing Centre (UNICC).

Layout and Design

Denian Ouyang
Associate Communications Officer (Multimedia)
United Nations International Computing Centre (UNICC).

Acknowledgements

This initiative was made possible through the collaboration of dedicated individuals and institutions committed to ethical innovation in digital transformation.
UNICC extends its appreciation to:

- **New York University – School of Professional Studies (NYU SPS)** for institutional partnership and academic excellence in shaping the capstone experience.
- **Professor Andres Fortino** and **Siranush Kostanyan** for academic leadership, mentorship, and the design of this high-impact collaborative framework.
- **The Product Management Team**, whose leadership in Agile coaching and stakeholder coordination was instrumental to the project's success.
- **Participating NYU Graduate Students**, from the Department of Business in the School of Professional Studies, whose creativity, professionalism, and ethical commitment transformed ideas into impactful technological solutions.

Required citation: UNICC, NYUSPS. 2025. *Responsive AI Innovation for Social Impact: A Case Study in Multilingual Media Moderation.*

Executive Summary

Key Elements and Strategic Findings

Key Element	Summary & Strategic Findings
Challenge Addressed	Misinformation, xenophobia, and declining media ethics pose threats to inclusive discourse. Social media platforms have reduced moderation, amplifying the spread of harmful content. Simultaneously, academic curricula often lag behind the real-world demands of AI-driven ethical media analysis.
Partnership Innovation	A collaboration between UNICC and NYU SPS used Generative AI and Agile methodologies to develop a multilingual, multimodal prototype for media analysis. The partnership combined student-driven capstone work with real-world UN goals.
Technological Solution	AI-powered tool composed of modular components: xenophobia detection, sentiment analysis, topic-based harm classification, and an interactive dashboard. Built to process media content in six UN languages and across text, audio, and video formats.
Educational Impact	Students gained applied experience with NLP, ethical AI, and Agile development, improving career readiness and producing publishable-grade work. Faculty noted higher engagement and quality compared to non-AI capstones.
Ethical Integration	Ethical risk reflection was embedded as a core deliverable, not an afterthought. Students addressed bias, cultural context, and fairness in model design—an innovation in AI education.
Agile Implementation	A dedicated Product Manager guided student teams using industry-standard Agile practices, including sprints, retrospectives, and product demos. This enabled complex, real-world project management within a three-month academic cycle.
Results	A functional prototype with four integrated AI modules, detailed ethical documentation, stakeholder engagement artifacts, and open-source code repositories. UNICC praised the tool's future deployment potential and alignment with UN digital priorities.
Recommendations	Scale the model globally via academic–UN partnerships; institutionalize Agile and ethical AI frameworks across UN projects; formalize Product Manager roles; develop a reusable AI ethics education toolkit; and consider operationalizing the prototype in human rights and crisis response contexts/

1 Introduction

1.1 Background on UNICC's role and mission

The United Nations International Computing Centre (UNICC) is the largest strategic partner for common, trusted, and cybersecure digital foundations across the entire UN system. Through state-of-the-art ICT infrastructure, digital tools, cybersecurity, cloud, data and artificial intelligence solutions, UNICC promotes progress toward a more connected, secure, and sustainable UN. With over 50 years of experience, UNICC supports the digital transformation and future of the UN family and other international organizations, delivering scalable and innovative solutions through a cost-effective and shared services model. This approach empowers our partners and clients to accelerate the adoption of required technologies to better serve global needs.

In line with this mission, UNICC collaborated with the New York University School of Professional Studies (NYU SPS) to pilot an educational innovation aimed at addressing one of today's most urgent challenges: the ethical use of artificial intelligence (AI) in digital media. This partnership exemplifies UNICC's commitment to capacity-building and responsible innovation, leveraging AI and Agile methodologies to bridge gaps between academic curricula and real-world digital governance needs.

1.2 Project Origin and Objectives

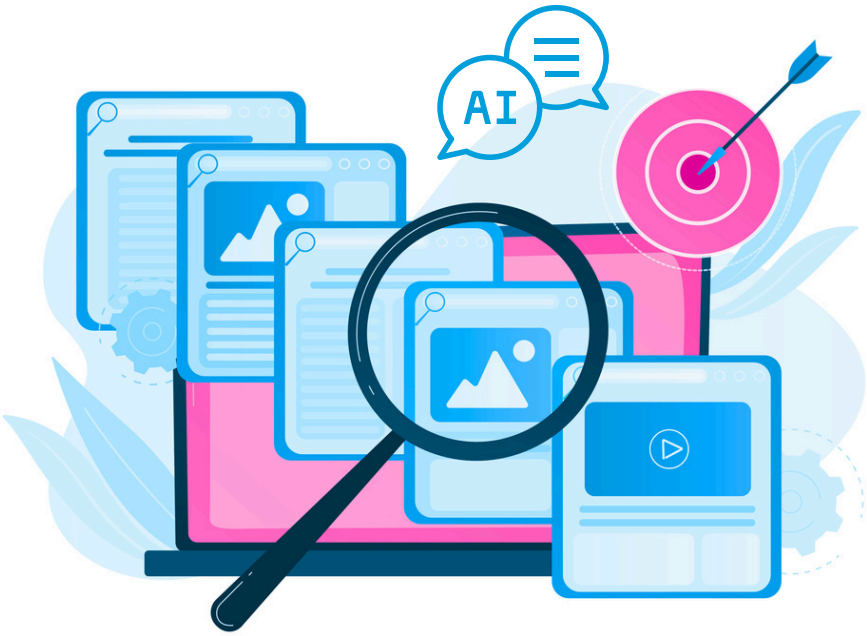
This initiative was catalyzed by a shared recognition of two interlinked issues: the escalating spread of harmful narratives and misinformation in digital media, and the growing skills gap between traditional educational models and the demands of an AI-driven economy. In response, UNICC sponsored a collaborative project with the New York University (NYU) School of Professional Studies to develop a prototype media analysis tool powered by artificial intelligence—designed to identify xenophobic language and other forms of media harm—while embedding ethical considerations and Agile development principles throughout its lifecycle.

The project aimed not only to develop an AI-driven tool aligned with UN values but also to create a replicable academic framework for preparing students to become ethical, innovation-ready professionals. Capstone projects within the program were designed to simulate real-world digital innovation environments, where students applied AI technologies, Agile frameworks, and stakeholder collaboration to address complex global challenges.

1.3 Addressing the Urgency of Ethical application of AI to Media Content Moderation

The proliferation of misinformation and bias in digital platforms stemming from the rise of social networks disproportionately affects vulnerable communities and threatens inclusive, fact-based discourse. This has been made considerably worse by the recent reduction in content moderation by social media companies. Xenophobia and toxic, harmful narratives often go undetected or unchallenged due to the lack of culturally sensitive, scalable detection tools. UNICC's sponsorship of this initiative reflects a broader institutional commitment to digital equity, human dignity, and the responsible use of AI.

Concurrently, the project recognized the need to transform how educational institutions prepare future technologists. It underscored the imperative of embedding ethics, stakeholder engagement, and iterative development practices into technical training. By combining technological innovation with mission-driven design, this initiative aligns with UNICC's vision of a more inclusive and just digital future, powered by partnerships and grounded in shared values.



Project Context and Rationale

2.1 Defining the Challenge

In an era marked by rapid technological advancement and the widespread dissemination of information through digital platforms, the rise of misinformation, xenophobia, and exclusionary narratives has become a critical concern. These trends pose tangible threats to social cohesion, public trust, and the safety of marginalized communities. Despite the availability of technological solutions, existing systems often lack the contextual awareness and cultural sensitivity required to detect and mitigate such content effectively. The drastic scaling back of content moderation by social media companies has compounded the problem.

UNICC identified this gap as a pressing issue aligned with its mandate to foster responsible and inclusive digital transformation across the UN system. At the same time, the organization recognized a parallel challenge within the education sector: the growing disconnect between traditional academic curricula and the evolving demands of the digital economy. Students and future professionals require hands-on, ethically grounded experiences that prepare them to navigate complex real-world challenges using emerging technologies.

This dual challenge—combating harmful narratives while strengthening digital literacy and capacity—formed the foundation for the UNICC–NYU collaboration.

2.2 Stakeholders and Strategic Value

This initiative was designed to serve a diverse and interconnected group of stakeholders, each with a distinct role in advancing ethical and effective media monitoring:

- **UNICC**, as the sponsoring UN entity, sought a scalable, AI-powered solution to identify harmful digital narratives and support broader efforts in digital inclusion and media ethics.

- **Students** benefited from direct exposure to emerging technologies and real-world applications of artificial intelligence, gaining the practical and ethical competencies essential for future roles in data science, digital governance, and international development.

- **Faculty and Academic Institutions** were provided with a scalable model for embedding experiential learning into STEM and project management curricula—modernizing teaching practices without compromising academic integrity.

- **The United Nations system and the global public** stand to gain from enhanced media transparency, stronger detection mechanisms for harmful content, and a pipeline of professionals equipped to uphold human rights in digital environments.

UNICC's leadership in convening this ecosystem of partners reflects its unique ability to transform technology into a force for inclusion, accountability, and innovation.

2.3 Alignment with Sustainable Development Goals (SDGs)

The project aligns closely with the United Nations' broader commitment to inclusive digital development and directly supports several Sustainable Development Goals (SDGs), including:

- **SDG 4 – Quality Education:** By introducing experiential, ethical, and future-oriented learning models.

- **SDG 5 and SDG 10 – Gender Equality and Reduced Inequalities:** Through the development of tools designed to expose and mitigate bias and discrimination in digital media.

- **SDG 16 – Peace, Justice and Strong Institutions:** By fostering tools that improve transparency, counter misinformation, and strengthen the digital information environment.

- **SDG 17 – Partnerships for the Goals:** By demonstrating how cross-sector collaboration can catalyze innovation and capacity-building.

Through this initiative, UNICC reaffirmed its role as a digital enabler for the UN system—translating global development goals into actionable solutions that promote fairness, inclusion, and resilience in the digital age.

Objectives and Scope of the Project

3.1 Strategic Objectives

This initiative was conceived as a strategic partnership between the United Nations International Computing Centre (UNICC) and New York University's School of Professional Studies (NYU SPS), with the dual aim of addressing an urgent global need and building future-ready capacity in the next generation of technology professionals.

The central objective was the design and development of an AI-powered media analysis prototype, capable of identifying xenophobic, harmful, and biased narratives across digital content. This tool would support UNICC's commitment to advancing responsible, rights-based approaches to digital innovation, while offering students a high-impact learning experience grounded in Agile development and ethical AI practices.

More specifically, the project sought to:

- **Develop a modular, AI-driven media tool** capable of detecting xenophobic and harmful language in news articles, podcasts, and social media posts.
- **Equip graduate students** with technical expertise in natural language processing (NLP), machine learning (ML), and prompt engineering within a real-world, mission-driven context.
- **Embed Agile methodologies** into academic workflows, promoting iterative development, stakeholder engagement, and cross-functional collaboration.
- **Integrate ethical design frameworks,** encouraging students to consider fairness, bias mitigation, and social responsibility throughout the AI development lifecycle.
- **Create a replicable Academia-UN partnership model** that bridges theory and practice, and contributes to the United Nations' broader digital innovation agenda.

3.2 Scope of Work and Implementation Parameters

Due to the nature of the student project process flows (final semester capstone, individual work, 300-hour project), the totality of the desired deliverables had to be broken up into two semester phases. The multimodal multilingual content moderation tool was broken up into components that could be tackled as individual components and spread over a two-semester cycle.

- **Phase 1:** First semester. A Proof of Concept prototype for text-based media that tackles detecting toxicity and xenophobia based on topics with a usable interface. The project was accomplished by a team of four students. These served as a foundation for technical skill-building and agile familiarity.

- **Phase 2:** Second Semester. Development of the final prototype that could manage moderation of content in the six UN languages and in three media modes: text, podcasts and videocasts.

Both phases required coordinated team-based efforts, supported by a dedicated Product Manager, focused on building a unified AI-powered media analysis prototype for UNICC. Capstone teams each developed a core module—xenophobia detection, sentiment analysis, topic classification, and front-end integration—within a structured Agile framework.

The project was organized as a competition between student teams during each semester. The Phase 1 projects attracted 16 students, formed into four teams. One team was selected as the first-round winner and went on to present their work to UNICC staff and colleagues around the world. The second round, Phase 2, garnered 64 applications formed into 16 teams. Again, the winner of the competition presented their work at the UNICC.

3.3 Scope of Innovation

- The tool analyzed text-based media (articles, tweets, podcasts) in English.
- It was architected with a modular design to allow future integration of new languages, formats (e.g., video, audio), and advanced detection layers.
- The prototype featured a user-friendly dashboard, a real-time testing interface, and documented API endpoints for integration.
- Organizing the students into a competition yielded a plethora of solutions for the UNICC technical teams to choose to follow up and implement.

3.4 Key Limitations

- The academic timeframe (three months) limited the depth of testing, scalability, and deployment.
- The tool remained at the prototype stage and was not launched operationally during the pilot.
- Formal AI bias audits and independent ethical review processes were outside the project's immediate scope but remain essential for future phases.

Despite these constraints, the initiative successfully demonstrated the feasibility and value of leveraging Academia-UN collaborations to generate ethically aligned, AI-driven innovations that serve public interest missions.

Methodology and Approach

4.1 Technological Foundation

At the core of this initiative was the application of advanced AI technologies and Agile project management frameworks—two mutually reinforcing pillars central to UNICC's digital transformation mandate. By combining the analytical power of Generative AI with the adaptability of Agile delivery models, the project created an innovation environment that was both technically robust and pedagogically effective.

The AI components included:

- **Large Language Models (LLMs)** such as GPT-4 and BERT for natural language understanding and semantic pattern detection.
- **Prompt engineering techniques** to tailor AI responses for detecting xenophobia, sentiment, and harm classification.
- **Python and R environments** for code development, data preprocessing, and model evaluation.
- **No-code AI tools** to support inclusive participation across student teams with varying levels of technical expertise.

The Agile project management approach featured:

- Sprint-based development cycles, allowing for time-bound iterations aligned with academic schedules.
- Structured backlog refinement, planning meetings, and retrospectives.
- Continuous integration of stakeholder feedback, including regular validation sessions with UNICC.
- A dedicated **Product Manager**, who served as the conduit between UNICC, student teams, and faculty mentors.

This methodological structure reflected industry best practices and mirrored the delivery environments found in digital innovation teams across the UN system.

4.2 Project Phases, from Design to Implementation

The initiative was implemented in two integrated phases:

Phase 1: Skill-Building through Individual Capstone Projects
Graduate students selected real-world challenges involving generative AI, including:

- Chatbot development for legal and labor applications
- Automated assessments of task feasibility
- Text mining for educational content generation.

These capstones allowed students to explore ethical and technical complexities while gaining hands-on experience in deploying AI tools. Deliverables included functioning prototypes, user documentation, and academic presentations assessed by both faculty and external experts.

Phase 2: Development of UNICC's AI Media Analysis Tool
This phase brought together multiple capstone teams under a shared challenge: to design an AI tool capable of:

- Detecting xenophobic and exclusionary language,
- Performing sentiment analysis across diverse media formats,
- Classifying content based on topic and harm potential,
- Integrating components into a unified, interactive dashboard.

The Product Manager, working closely with UNICC and NYU faculty, decomposed the complex challenge into four distinct development tracks—each aligned with a specific functionality. Agile practices facilitated concurrent development, iterative refinement, and technical integration, resulting in a cohesive prototype.

4.3 Partnership Model and Roles

The initiative's success stemmed from a multi-stakeholder collaboration that embodied the UN's spirit of innovation through cooperation:

- **UNICC** provided the strategic challenge, served as the end user, and guided ethical and technical standards aligned with UN mandates.
- **NYU SPS** contributed academic structure, curriculum integration, and faculty mentorship.
- **Students**, many of whom were professionals with industry experience, formed cross-functional teams that mirrored real-world product development settings.
- **The Product Manager** played a central coordinating role—ensuring Agile discipline, facilitating stakeholder alignment, and maintaining momentum across all tracks.

This collaborative structure not only ensured alignment with institutional values but also demonstrated the potential of academic-UN partnerships as vehicles for both innovation and capacity-building.

Challenges Encountered and Lessons Learned

5.1 Navigating Complexity in Applied AI and Academic Collaboration

Implementing a mission-driven AI prototype within a graduate-level academic structure presented several technical, operational, and contextual challenges. These obstacles offered valuable insights for UNICC and its academic partners in shaping future innovation engagements.

Theme	Challenge	Lesson Learned
Balancing Real-World Complexity	Scope exceeded typical academic limits (NLP, ethical AI, agile dev)	Effective project decomposition and modular design allow ambitious real-world problems to be tackled within academic settings.
Introducing Agile in Classroom	Students had limited Agile exposure; cultural shift required	A dedicated Product Manager (separate from faculty) was key to embedding Agile discipline and ensuring team accountability.
Managing Cross-Team Dependencies	Modular design introduced integration risks due to interdependencies	Standardized documentation, version control, and cross-team syncs mitigated integration issues and delays.
Addressing Ethical Risk in AI	Designing for xenophobia detection involved bias, cultural context, and misclassification risks	Making ethical reflection a core deliverable helped students develop responsible AI aligned with UN values.
Operating Within Tight Timeframes	Semester length constrained iteration and technical depth	Agile's incremental approach enabled delivery within deadlines; longer projects may benefit from phased, multi-semester planning.

5.2 Strategies for mitigation and adaptation

In response to these challenges, the project introduced several practices that contributed to its success and can serve as models for future UNICC–Academic collaborations:

- **Dedicated Product Manager Role:** Bridged academic and operational expectations, facilitated Agile discipline, and ensured alignment with UNICC's technical and ethical goals.
- **Embedded Performance Metrics (OKRs and KPIs):** Provided a shared structure for evaluation, accountability, and cross-team consistency.
- **Hybrid Collaboration Model:** Blended team competition with shared responsibility to foster innovation, ownership, and high performance.
- **Stakeholder Co-Design:** Active involvement from UNICC throughout the project lifecycle ensured relevance, enhanced quality, and strengthened the students' understanding of UN priorities.

Through these strategies, the initiative not only overcame operational barriers but also established a foundation for future collaborations that bridge institutional, academic, and developmental goals.

6 Impacts and Results

6.1 Tangible outcomes and qualitative/quantitative results

The UNICC–NYU collaboration resulted in a range of high-impact deliverables that advanced both educational and institutional goals. These outcomes demonstrate the potential of cross-sector innovation to produce scalable, ethically grounded digital solutions.

Key deliverables included:

- **Four Functional AI Modules**, each aligned with a specific area of media analysis:
  - Xenophobic language detection
  - Sentiment analysis
  - Topic-based harm classification
  - Integration and front-end dashboard

- **A Unified Prototype Tool** capable of processing diverse digital media formats, such as articles, tweets, and podcast transcripts, into actionable insights aligned with UNICC's mission of promoting ethical digital transformation.

- **Ethical AI Design Documentation**, including fairness and bias considerations, transparent development logs, and guidance for future evaluation, reinforcing UN values of dignity, inclusion, and non-discrimination.

- **Agile Project Artefacts**, such as sprint logs, burndown charts, and retrospectives, validate students' application of industry-standard project management practices.

- **Open-Source Repositories and Draft Research Outputs**, including GitHub documentation and academic write-ups, establish a knowledge base for future adaptation, peer learning, and publication.

These outcomes offer a replicable model for future UNICC engagements that integrate education, capacity-building, and technology co-creation.

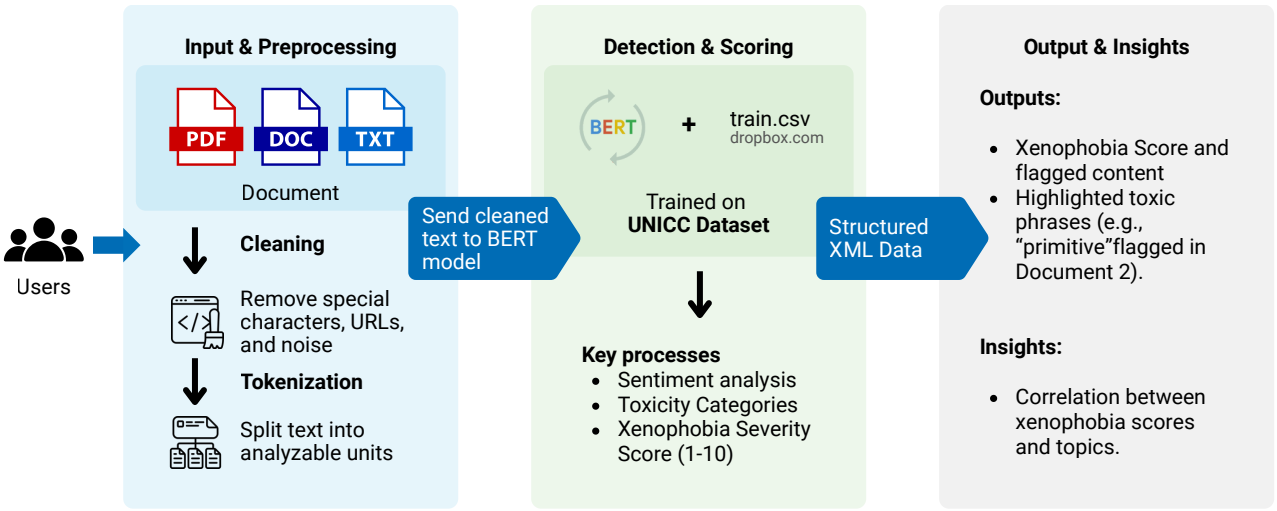
6.2 Phase 1 – Proof of Concept for Text-Based Model

Model Construction

Following the development of the Phase 1 model, the winning capstone team created a modular AI system composed of three integrated analytic components: xenophobic language detection, fact and language validation, and topic-based narrative analysis. Together, these modules formed the foundation of the Phase 1 prototype and were unified by an interactive dashboard and an API-based backend, allowing for seamless real-time media analysis. The modular architecture ensured that each function could operate independently while still contributing to a cohesive user experience. Figures 1 through 3 illustrate the mechanics and interface of each module.

Figure 1 presents the xenophobic language detection module, the primary engine for identifying harmful language in media content. This module was built using a fine-tuned BERT model and processes text extracted from various document formats such as Word files, PDFs, and plain text. After cleaning and tokenizing the input, the model analyzes the data for both overt and covert xenophobic content. Identified phrases are scored on a severity scale from 1 to 10, and each is accompanied by a contextual explanation. Results are output in XML format and rendered in real-time on the dashboard. The interface is designed for clarity and immediate action, enabling media professionals to quickly detect and address problematic language through an intuitive visualization of flagged phrases and severity scores.

Figure 1: Xenophobic Language Detection Module Dashboard Displaying Real-Time Analysis Results Including Severity Scoring and Content Flagging



In Figure 2, the fact and language validation module is shown. This module enriches the tool's ethical oversight by assessing whether the language used is both factually accurate and free from stigmatizing or biased expressions. It compares input text against a curated glossary of migration-related terminology, including authoritative sources such as UNHCR. It also uses sentiment-aware algorithms to highlight emotionally charged phrases—such as "flood of immigrants"—and offers neutral alternatives tailored to the context. The dashboard interface displays original phrases side-by-side with suggested revisions, supporting informed editorial decision-making and encouraging a shift toward inclusive and respectful reporting.

Figure 2: Fact and Language Validation Module Interface Showing Term Cross-Referencing and Alternative Language Suggestions

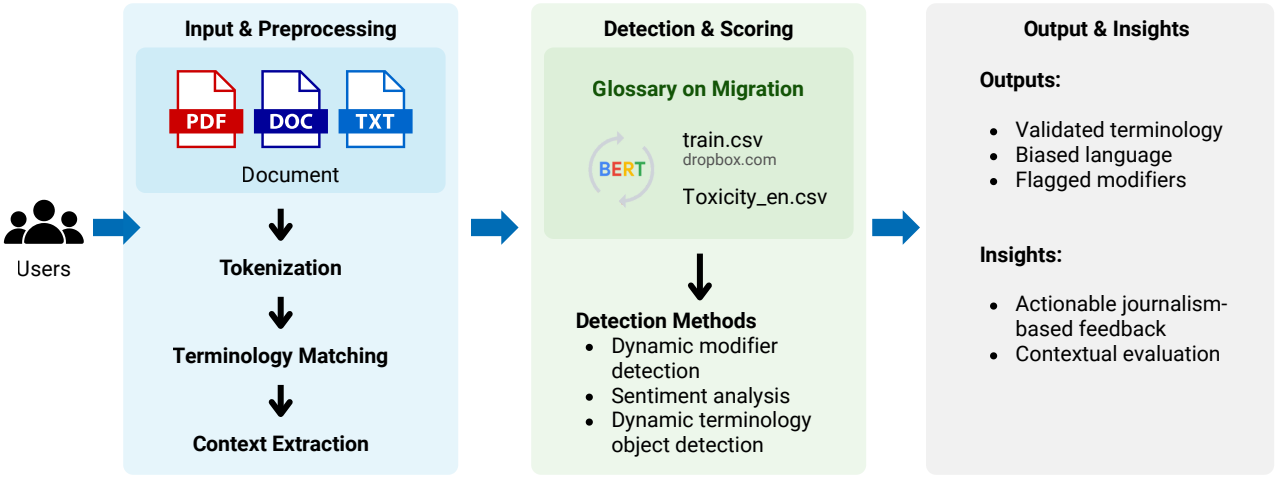
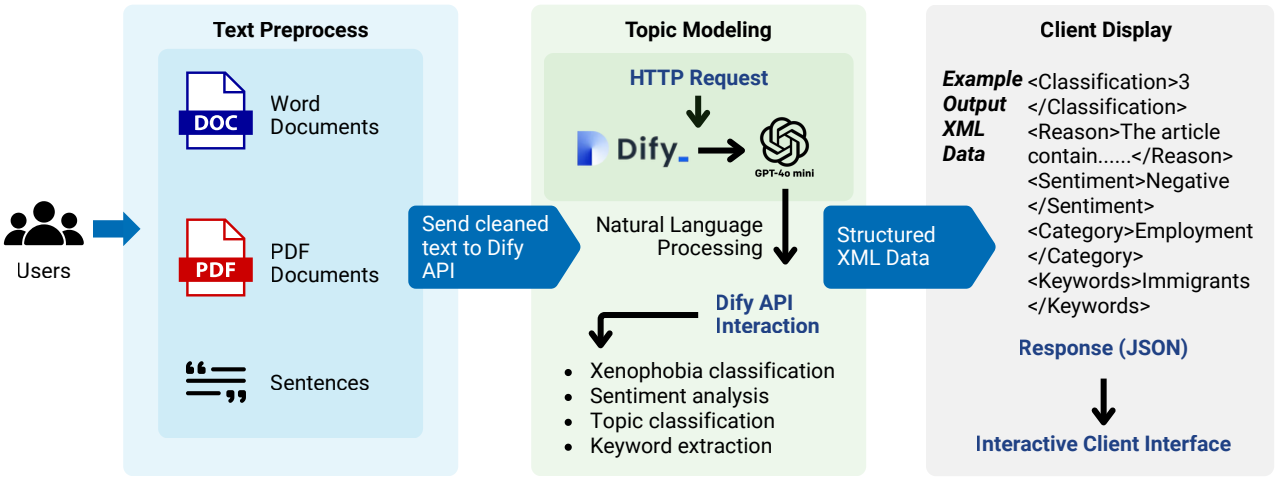


Figure 3 illustrates the topic-based narrative analysis module, which offers a broader understanding of the thematic and emotional structure of the analyzed text. Using a fine-tuned GPT-4 model accessed through the Dify API, this module performs topic modeling and classifies content under thematic headings like "Employment," "Public Safety," or "Social Contributions." Each theme is assigned a sentiment score—positive, neutral, or negative—providing users with insight into the emotional framing of the narrative. Dominant keywords are also extracted to highlight recurring ideas or rhetorical patterns. The visual interface allows users to explore both thematic distribution and sentiment at a glance, supporting editorial teams in recognizing patterns and making more intentional content choices.

Figure 3: Topic-Based Analysis Dashboard Presenting Thematic Classification and Sentiment Analysis Results



System integration was accomplished through an API-based architecture that combined the three modules into a single interactive platform. Each component's output was structured in XML and fed into a unified dashboard, enabling cross-module consistency. A chatbot interface supported dynamic engagement, allowing users to query flagged content and receive real-time feedback. This integration ensured that the tool not only performed its functions independently but also delivered a seamless, comprehensive analysis experience for end users. Together, the modules and their integration represent a technically robust and ethically attuned solution that demonstrates the feasibility of AI-driven content analysis in support of inclusive, rights-based.

Model Validation

To evaluate the effectiveness of the prototype, the development team conducted a systematic accuracy assessment using a labeled dataset of 159,000 entries, which included xenophobic, toxic, and neutral content sourced from migration-related news articles, social media posts, and opinion pieces. The dataset was carefully annotated by experts and preprocessed through normalization, tokenization, and text cleaning. The model was evaluated using standard metrics—precision, recall, and F1 score—to assess its ability to detect both explicit and implicit xenophobic language. Real-time performance was also assessed, with a target response time of under two seconds per input. The results demonstrated a precision of 91%, recall of 87%, and an F1 score of 89%, indicating strong performance in accurately identifying harmful content while minimizing false positives and negatives. The average severity score assigned to toxic content was 7.8 out of 10, and review times were reduced by 60%, confirming the tool's potential to streamline editorial workflows. User feedback from media professionals validated the model's usability, particularly its effectiveness in surfacing subtle forms of bias such as coded language and emotional framing.

6.3 Phase 2 – Extension of Phase 1 to Multimodel and Multilanguage Models

Model Construction

The MIRROR system was developed as a Phase 2 model for the UNICC project, aiming to establish a unified platform for the detection of misleading information, xenophobic, and harmful content across all media forms. Its core purpose is to provide ethical content detection and multilingual processing through a single drag and drop interface, designed so that users do not need to click to upload files.

The system's operational flow, as depicted in the User Journey and System Flow Chart diagrams, begins with a user accessing the platform via the internet and logging in through an authentication process. Users are presented with a landing page where they can drag and drop any type of file, including text, audio, or video. After validation and storage, the system processes the file. For audio and video, the Analysis Engine performs Automatic Speech Recognition (ASR) for transcripts and frame extraction for labeled objects. If the content is not in English, the transcript undergoes multilingual analysis. The resulting textual data is then fed into the Phase I modules for analysis of harmful content or misinformation. A key step involves calling a fact-checking API to check the worthiness of any key claims made in the source material. Finally, the system generates and displays a report detailing findings such as noted segments, explanations, potential policy risks, and fact-checking worthiness. Users can also access their history of reports through a portal.

Technically, the system is built entirely on the Vercel platform using modern web tools for speed and smooth navigation. It leverages the Vercel AI SDK to process files dropped by the user. Large Language Models (LLMs) are central to the system for efficient multimodal processing due to their natural advantage with human-created content. Specifically, Gemini 2.5 from Google is used to generalize responses, chosen for its ease of use. Recognizing that LLM output accuracy "never hits 100%", a fact-checking system was implemented to mitigate this risk and enhance reliability. This API, referred to as the "clam booster Api" and obtained from a PhD team at the University of Texas and Arlington, checks claims and offers verification suggestions rather than definitive answers. A modern frontend framework was utilized to improve the user experience, and built-in authentication ensures secure login.

The development process involved several challenges. An early decision was made not to leverage the Phase 1 work as it was considered difficult to integrate and had limitations in scalability and integration potential with the new platform. Selecting the appropriate AI model for analyzing various media types was another hurdle; the team explored Azure AI video and index analyzer and Vertex AI from Google before choosing Gemini due to its ease of use, despite it being a new beta platform. Ensuring the trustworthiness of the LLM output directly led to

the integration of the fact-checking system. Multilingual processing presented difficulties. While the system supports languages like Arabic, Chinese, English, French, Russian, and Spanish for analysis, the fact-checking API only functions in English . Therefore, content in other languages must be translated to English before being sent to the fact-checking API. Experimentation with different LLM prompts was conducted to improve accuracy across various languages. Other technical issues included a production problem with file size limits; the production system was limited to 4.5 MB, whereas the target was 200 MB. Future improvements include enhancing frame extraction from videos for greater accuracy and scaling the solution for growing media volume and a larger number of users.

A systematic testing method and evaluation framework were developed to ensure the system's completeness and accuracy. The framework assesses accuracy and reliability by assigning a toxicity grade score on a scale of 0 to 100 and classifying results into categories: non-toxic (0-49), mild (50-69), high (70-89), and extreme high toxicity (90-100). Performance levels (high, medium, low standard) were defined based on how much the system's score deviates from manually labeled ground truth and the percentage of correct category assignments. Testing was conducted using three types of data, as presented in Appendix D: documents, social media records, and public resources from platforms like TikTok, YouTube, and X. The use of public resource content was specifically chosen to validate the model's generalization ability in less controlled, noisy, real-world environments and simulate user scenarios. The team hand-coded or manually labeled this public data themselves for validation purposes.

Model Validation

To ensure the system's completeness and accuracy, a systematic testing method and evaluation framework were developed. This framework assesses accuracy and reliability by assigning a toxicity grade score on a scale of 0 to 100 and classifying results into categories: non-toxic (0-49), mild (50-69), high (70-89), and extreme high toxicity (90-100). Performance levels (high, medium, low standard) were defined based on how much the system's score deviates from manually labeled ground truth and the percentage of correct category assignments, aiming for at least 95% consistency for high standard. Testing was conducted using three types of data, as presented in Appendix D: documents, social media records, and public resources from platforms like TikTok, YouTube, and X. The use of public resource content was specifically chosen to validate the model's generalization ability in less controlled, noisy, real-world environments and simulate user scenarios. This public data was hand-coded or manually labeled by the team themselves for validation. The test results showed an overall accuracy of 87.18% for documents, 92.03% for social media records, and an impressive 98.29% for public content across text, audio, and video data testing. Specifically, accuracies were 92.12% for text, 85.32% for audio, and 85.10% for video in documents; 92.03% for text in social media (audio and video were N/A); and 99.41% for text, 99.30% for audio, and 98.17% for video in public resources. These results demonstrate the system's capability for effectively detecting toxic content at scale in real-world scenarios.

6.4 Over Quantitative Results for Team-Based Projects

An analysis of 111 capstone projects over four years—including those before and after AI and Agile integration—demonstrated significant positive shifts in performance:

- **AI-based projects** achieved:
  - 45.3% "A" client satisfaction ratings vs. 29.8% for non-AI projects
  - 26.6% publication-worthiness vs. 6.4% for non-AI projects
    (Statistically significant at $p < 0.01$)

- **Team-based projects** had a higher rate of "A" satisfaction ratings (68.8%) compared to individual efforts (33.7%)
  ($p < 0.05$)

- **Projects with a Product Manager** outperformed those without:
  - 54.8% "A" satisfaction vs. 32.5% without PM
    ($p < 0.05$)

These findings support the hypothesis that structured, interdisciplinary, and Agile-aligned approaches yield higher-impact deliverables, particularly in development contexts involving ethical technology and stakeholder engagement.

6.5 Stakeholder feedback and engagement outcomes

Feedback from stakeholders underscored the initiative's relevance, quality, and strategic alignment:

- **UNICC** representatives expressed strong support for the prototype's functional and ethical design, emphasizing its potential as a future tool in digital media governance and early warning systems.

- **Students** reported significant growth in AI literacy, ethical reasoning, and professional readiness. Many credited the experience with securing job offers or advancing careers in technology and public service.

- **Faculty mentors** observed increased engagement, more sophisticated deliverables, and deeper learning outcomes when compared with previous non-AI capstone cycles.

Beyond outputs, the project instilled a sense of civic responsibility and ethical purpose in participants—further reinforcing UNICC's commitment to building not only digital tools, but also principled digital leaders.

7 Recommendations and Future Directions

7.1 Recommendations

Drawing on the outcomes of this pilot, UNICC proposes the following strategic actions to scale the model, strengthen institutional innovation pipelines, and deepen the impact of ethical AI across the UN system and beyond:

- **Institutionalize Agile Practices**

  Encourage the broader adoption of Agile methodologies—especially in pilot-stage or youth-engagement initiatives. Structured Agile frameworks accelerate delivery, promote stakeholder alignment, and support learning through iteration.

- **Scalability and adaptability to other contexts**

  The modular design of the AI media analysis tool positions it for adaptation across multiple UN programmatic areas. Future use cases could include:
  - **Human Rights Monitoring**: Detecting discrimination or hate speech in online discourse.
  - **Crisis Response and Peacebuilding**: Identifying harmful narratives during humanitarian emergencies.
  - **Gender and Social Inclusion**: Evaluating representation of marginalized groups in media content.
  - **Conflict Early Warning**: Monitoring sentiment and misinformation in fragile or post-conflict settings.

In academic settings, the model is equally adaptable to fields beyond data science—such as journalism, public policy, cybersecurity, and global affairs—promoting interdisciplinary collaboration on digital governance challenges.

7.2 Suggested future initiatives.

To build on this successful pilot, UNICC recommends the following next-step initiatives:

- **Phase 3: Development and Deployment**

  Deploy the current prototype to UN staff to support:
  - Multilingual analysis (across all six UN official languages),
  - Multi-modal formats (including video and audio),
  - Operational readiness for deployment in field or HQ contexts.

- **Ethical AI Education Toolkit**

  Develop and distribute a training package for universities, based on this model. Topics to include bias mitigation, prompt engineering, stakeholder-centered design, and UN-aligned AI governance.

Appendices

Glossary, technical diagrams, and detailed data

Appendix A:

Glossary of Key Terms

Term	Definition
Agile Methodology	A project management approach based on iterative development, where solutions evolve through collaboration between cross-functional teams.
Artificial Intelligence (AI)	The simulation of human intelligence in machines, enabling them to perform tasks such as learning, reasoning, and decision-making.
Generative AI	A subfield of AI that uses models like GPT-4 to generate new content, including text, images, or code, based on learned patterns.
Large Language Model (LLM)	A type of AI model trained on massive text datasets to understand and generate human-like language.
Prompt Engineering	The process of crafting inputs (prompts) to guide AI models in producing desired outputs.
Xenophobia Detection	The use of AI tools to identify language or narratives that express prejudice against people from other countries or cultures.
Sprint	A fixed period of time (typically 1–2 weeks) during which specific project tasks are completed in Agile frameworks.
Product Manager	A role responsible for aligning project delivery with client expectations, coordinating between teams, and facilitating Agile processes.

Appendix B:

Capstone Project Roadmap

Week	Milesont
Week 1	Team formation and onboarding; scope definition
Week 2-3	Sprint 1: Functional requirement development
Week 4-5	Sprint 2: MVP (Minimum Viable Product) builds for individual components
Week 6-7	Sprint 3: Mid-point reviews and refinements based on stakeholder feedback
Week 8-9	Sprint 4: Integration testing and UI development
Week 10	Sprint 5: Final testing and documentation
Week 11	Client showcase presentations and tool handover
Week 12	Retrospective reflections and academic submission

Appendix C:

Evaluation Rubric and Criteria

- **Client Satisfaction (40%)**–Based on clarity, functionality, and relevance of the final product to UNICC's stated needs.
- **Technical Accuracy (25%)**–Measured by prototype performance, code quality, and usability.
- **Agile Process Compliance (20%)**–Assessed via sprint documentation, participation in retrospectives, and milestone achievement.
- **Ethical Considerations (15%)**–Inclusion of fairness, transparency, and bias-mitigation strategies.

Appendix D:

Creation and Intended Use of Synthetic Datasets for Model Accuracy and Validation Testing

To rigorously evaluate the AI media analysis tool developed through the UNICC–NYU collaboration, a synthetic dataset was specifically created for training, accuracy testing, and cross-modal validation of the prototype's detection capabilities. This dataset was designed to simulate real-world media content containing varying degrees of toxicity and xenophobic language while ensuring consistency, scalability, and multilingual reach. The dataset comprises synthetically generated news articles and social media posts as well as multimodal data (audio and video) with aligned transcripts, all developed in six official UN languages: English, Spanish, French, Russian, Chinese, and Arabic.

The text corpus was generated using GPT-4, starting with English-language stories and mechanically translating them into the other five languages using both ChatGPT and Google Translate. Each article was assigned one of five toxicity levels—None, Low, Medium, High, or Maximum—and both original and translated versions were labeled accordingly. For social media content, multilingual posts were hand-coded by researchers with binary toxicity labels and an additional indicator for extreme toxicity. This ensured a balanced representation of subtle, moderate, and overtly harmful expressions. The metadata for each file includes the toxicity level assigned during human coding.

In an innovative expansion, the dataset also included multimodal elements to test the models' performance across content formats. Using KreadoAI, synthetic video and audio files were produced from the base text. A consistent synthetic avatar—a male professional in a seated position—was used for all video presentations to eliminate visual bias. Audio recordings were created with synthetic native speakers in each language, selected for linguistic authenticity and tonal neutrality. The audio files were directly extracted from the videos, ensuring complete alignment of content across modalities. Each video and audio file was approximately two minutes in length, standardized to ensure consistency across languages and toxicity levels. Scripts (transcripts) in all six languages were also generated and verified for accuracy, serving as a reference for testing transcript-based NLP models.

The dataset's composition allowed for cross-lingual and cross-modal benchmarking of model performance. Students and evaluators were able to compute key metrics such as accuracy, precision, recall, and F1 score using both human-coded and machine-detected labels. Testing was conducted using a stratified subset of the dataset across different toxicity levels and languages, allowing fine-grained comparisons of the model's ability to detect harmful content in diverse and linguistically complex media formats. Real-time processing was also validated using this data, with performance thresholds set to reflect editorial use cases.

In addition to performance benchmarking, the dataset served as a critical component in validating the text extraction and classification pipeline. The included scripts allowed evaluators to compare raw video and audio outputs against text-based model interpretations, ensuring alignment and interpretability across modalities. Throughout, careful attention was given to ethical considerations, including content warnings for highly offensive material and guidance for responsible use, especially in educational or UN-affiliated environments.

This dataset, while synthetic, was meticulously designed to emulate the variability and complexity of real-world harmful narratives. Its multilingual, multimodal structure, rigorous labeling schema, and ethical safeguards made it a valuable asset for evaluating the AI prototype's effectiveness and generalizability. As such, it not only supported validation of the tool's initial deployment but also established a reusable framework for future enhancements and broader institutional testing.

School of Professional Studies
New York University

7 East 12th Street
New York, NY 10003
USA

www.sps.nyu.edu

UNICC AI Hub
United Nations International Computing Centre

Palais des Nations
1211 Geneva 10
Switzerland

www.unicc.org