MSc in Computer Science - Team Project

Project Plan

Project Title:

AI-Powered Social Media Threat Monitoring System for Protecting Freedom of Speech Advocates

Project Summary:

- What are you doing?

Recent tragedies, including public assassinations of individuals known for defending freedom of speech, highlight the growing risks faced by journalists, activists, and politicians. Online threats, harassment, and violent rhetoric often escalate from digital spaces to real-world harm. While commercial monitoring tools exist, they are primarily aimed at brands or governments, not at-risk individuals and NGOs. Our proposed project addresses this gap by developing an AI-driven system to monitor social media platforms for emerging threats against free speech advocates. The goal is to detect violent rhetoric, doxxing, and coordinated harassment campaigns in near real-time, helping to provide early warnings to those at risk. The project is defensive, proactive, and ethical, supporting safety without promoting surveillance or harm.

- Why are you doing it?

I have been inspired by the recent assassination of Charlie Kirk to develop an application that could reduce the likelihood of such incidents occurring in the future.

- Who will use it?

This project contributes a focused, open-source, user-oriented threat monitoring system tailored to individuals and NGOs advocating for freedom of speech.

- How will they use it? (Example use case)

Much like the Volatility Index (VIX) in the stock market, our application will continuously track social media posts and present results visually through charts and dashboards. For guest users, the software provides a general assessment of overall social media trends, indicating whether sentiment is calm, neutral, or radical. For registered users, the system allows configuration and saving of specific keywords so it can monitor mentions of particular individuals or organisations across platforms. If detected risks escalate beyond a defined threshold, the system will issue alerts via email or other notification methods. This enables journalists, activists, and NGOs to anticipate potential threats and make informed decisions about whether to proceed with events, strengthen communication strategies, or enhance security measures to safeguard personal safety.

Project Development:

A description of your first minimal viable product (MVP) including what you will learn from deploying this and what you would measure in order to learn it. (Even if you are not taking a lean approach to managing your project you should be able to articulate an MVP.)

- How will you build your system? (System diagram)
  - Front-end: User interface components
  - Back-end: Technical components
  - Data sources: What data will you use and how will you access it

System Architecture (High-level pipeline):

Data Ingestion → Preprocessing → Threat Classification (NLP) → Graph Analysis → Dashboard/Alerts

- Data Ingestion: Gather social media data from platforms such as Reddit, Bluesky, Mastodon, Facebook, and Instagram using their APIs or public datasets. Start with batch ingestion, add small-scale streaming as a stretch goal.
- Preprocessing: Tokenization, language detection (fastText), deduplication.
- Threat Classification: Fine-tune transformer models (BERT, RoBERTa) on hate speech datasets; compare with classical baselines.
- Graph Analysis (stretch goal): Use NetworkX or similar tools to analyze communities spreading threats.
- Dashboard: Svelte frontend + FastAPI backend for displaying alerts and trends.

For the datasets, we have decided to go with the following:

- https://pushshift.io/signup - This gives access to Reddit Api
- https://huggingface.co/collections/manueltonneau/hate-speech-supersets-664ef6d2bc40cce7a8b1092f - A multilingual hate speech dataset, eight languages are covered.
- https://huggingface.co/datasets/irlab-udc/metahate- A meta collection of thirty-six social media comments hate speech datasets

- https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech - A hate speech score dataset with ten labels

The methodology we decide to utilize for this specific project is SCRUM, for SCRUM roles product owner will prioritize features of the project, the SCRUM master will facilitate scrum events and track progress via burndown charts on GitHub projects, the development team will be cross-functional, and all three members shall be involved. A prioritized backlog of product features, sprint backlog with estimated tasks and incremental working prototypes. We have agreed upon four sprints that match our objectives, sprint one will cover week one to week three and will involve dataset ingestion, sprint two will cover week four to week seven and will involve model development, fine tuning the Deep Learning Models, sprint three encompasses week eight to week eleven and involves Evaluation of models, error analysis, initial results , Dashboard design and integration with model outputs. Sprint four covers week twelve to week fourteen and will involve: Graph analysis and system refinement, User testing, feedback collection, adjustments, Final report, documentation, and presentation. The tools used in this project will be GitHub projects as a project management tool, Google Colab for the models and Apache Spark.

Evaluation:

- How will you evaluate your system? (Initial ideas)
- How will the evaluation inform your future development?

The proposed system requires a rigorous approach to ensure accuracy, User-centric performance and ethical robustness. As a team we decided upon the following metrics to evaluate our proposed system:

1. Technical Metrics: Precision, recall, F1-score for classification; processing efficiency.
2. User-Centered Evaluation: Mock testing with peers role-playing journalists/NGOs.
3. Comparative Analysis: Benchmark against hate-speech datasets and existing tools.

The team's evaluation approach shall inform the future development of the proposed project by scheduling iterative improvements that will be synthesized into a backlog in the project management tool, this will allow high priority tasks to be on the top of the to-do list precedence and will also allow faster continuous improvements to take place.

MSc in Computer Science - Team Project

Project Plan

Project Management:

- How will you run the project?

Our team will use GitHub Projects as the primary project management tool. This will allow us to track tasks, assign responsibilities, and monitor progress in real time. We will adopt a Kanban approach, creating columns for To Do, In Progress, Review, and Done. Each task will be created as a GitHub Issue and linked to the project board. This ensures that code commits, pull requests, and progress updates are directly tied to visible tasks.

- What deadlines will you set?

We will follow the MSc weekly schedule for major milestones. The Project Plan will be submitted in Week 2. The Interim Demonstration Event and Interim Report are due in Week 6. The User/System Evaluation Presentation will take place in Week 9. The Final Demonstration Event is scheduled for Week 13. The Final Report is due in Week 14.

- What is success?

Success for our project will be measured in several ways. First, delivering a working prototype that detects and visualizes threats against free speech advocates demonstrates technical achievement. Second, achieving reliable evaluation metrics indicates research validity. Finally, positive feedback from peer evaluations will confirm usability. Together, these criteria will define both the academic and practical success of our project. In essence, if individuals can use this tool to anticipate hazards to some degree and thereby reduce their risk of sustaining physical harm in real-world scenarios, it would demonstrate the project's success.

MSc in Computer Science - Team Project

Project Plan

Team Name:

Zero Cool

Team Members:

Name	Student Number	Contact Number
Diwen Xiao	D24128462	+353 87 693 0108
Denis Muriuki	D22127693	+353899415276
Mingde Zhou	D24128243	+ 353 879330329

Team Meetings:

- Does everyone have to attend all meetings?

  Our team expects every member to attend all scheduled meetings unless there is an unavoidable conflict, in which case the person should inform the group in advance. This ensures transparency and accountability.

- How often will there be meetings?

  5 times a week.

- Are there topics that are out-of-bounds?

  The meeting will always focus on project topic and project progress, all team members agree to avoid talking about personal things to keep meetings productive, and decision-making will usually follow a majority rule approach, but for highly technical issues, we will rely on the expertise of the member most familiar with the topic. This balances fairness with efficiency.

- Online or face-to-face?

  Mostly online, sometimes face to face when we meet urgent issues, face to face meeting will be arranged in TUD campus.

- Decision making (Majority rule/Unanimous/Expertise wins)?

Most rules are suitable for most conditions, but when meeting highly technical issues, we will take opinions by who has the most experience.

- How will turn taking happen?

  Turn-taking during discussions will be managed by allowing each member the chance to speak in sequence, while ensuring no one dominates the conversation. This way, all voices are heard, and contributions are balanced.

Team Conflict:

- How do you deal with the habits of individual members?

  We respect different working habits, but members are expected to communicate openly and adjust if their habits affect group progress.

- How do you deal with unresolved issues?

  If ssues cannot be resolved immediately, they will be documented and revisited in the next meeting to ensure no concerns are ignored.

- How will you deal with conflict?

  Conflicts will be addressed directly in meetings, giving each member the opportunity to explain their perspective. Discussions will remain professional and focused on solutions.

- How will you avoid it?

  Clear task allocation, agreed deadlines, and regular communication will help minimize misunderstandings. Mutual respect and accountability will be emphasized from the beginning.

- Who will have the ultimate veto?

  In rare cases where consensus cannot be reached, the team will follow a majority rule system. However, if the issue is highly technical, the decision will be guided by the expertise of the member most familiar with the topic.