### **Capstone Project Concept Note and Implementation Plan**

### **Project Title: Civic Voice (Group 4)**

### 

### 

### 

### 

### 

### Group 4 Members Name:​​

### Alvin Diamond Johnson

### ​​​​ George Soe

### ​​​​​​​Samuel Aloysious Musa

### ​​​​​​​Charles P. Money

### ​​​​​​​ Louis L. Gweibeic

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Project Overview**

### The CivicVoice project is a digital platform designed to strengthen civic participation and improve public service delivery in Liberia. It directly addresses two critical **Sustainable Development Goals (SDGs)**:

### **SDG 16: Peace, Justice, and Strong Institutions**: The project aims to promote transparent, accountable, and inclusive governance by creating a direct channel for communication between citizens and institutions. It empowers citizens to report issues and enables authorities to respond efficiently, thereby building trust and strengthening institutional capacity.

### **SDG 11: Sustainable Cities and Communities**: By providing real-time data on urban and community-level challenges, CivicVoice helps local governments and NGOs identify and address problems more effectively. This data-driven approach supports more efficient resource allocation for infrastructure, security, and healthcare, contributing to the development of sustainable and resilient communities.

### **Problem and Solution**

The core problem we aim to address is the **communication gap between citizens and the government** in Liberia. Citizens often lack a reliable and accessible way to report community issues, leading to public service inefficiencies, resource misallocation, and a breakdown of trust. This is compounded by a reliance on manual, slow, and non-transparent reporting systems.

The CivicVoice app solves this problem by providing an integrated, AI-powered platform for citizen reporting. The solution's potential impact is significant:

* **For citizens**: It offers a user-friendly way to report issues via text or voice, ensuring their voices are heard and acknowledged. This fosters a sense of empowerment and civic duty.
* **For government and NGOs**: It transforms unstructured citizen feedback into actionable insights. By using **Natural Language Processing (NLP)** and **Geospatial Data Analysis**, the platform can automatically categorize issues, detect sentiment, and identify real-time "hotspots" of public concern. This allows for a proactive and data-driven approach to governance and resource allocation, leading to more efficient and targeted public services.

Ultimately, CivicVoice bridges the gap between digital technology and governance, creating a more responsive and accountable system that benefits the entire community.

### 

### **Project Objectives**

### The primary objective of the CivicVoice project is to develop and deploy an AI-powered platform that empowers citizens and provides actionable insights for government and non-governmental organizations (NGOs). To achieve this, we have outlined several specific objectives.

### **Key Objectives**

### Develop an Accessible Reporting Interface: Create a user-friendly mobile and web interface that allows citizens to easily submit reports on community issues via text or voice. This objective is crucial for ensuring the platform is inclusive and widely adopted, addressing the current inaccessibility of reporting systems.

### Implement an AI-Driven Classification Engine: Build and train machine learning models to automatically classify citizen reports into key categories such as education, healthcare, and infrastructure. This directly addresses the problem of manual and inefficient reporting systems by transforming unstructured data into organized, actionable information.

### Perform Real-time Sentiment Analysis: Develop a Natural Language Processing (NLP) pipeline to analyze the sentiment of citizen feedback. This will allow government and NGO officials to quickly gauge public opinion, identify areas of high dissatisfaction, and respond more proactively.

### 

### 

### 

### Visualize Geospatial Issue Hotspots: Integrate a geospatial data system to map the locations of reported incidents. This will provide a clear, visual understanding of where problems are concentrated, helping to optimize resource allocation and target interventions effectively.

### Generate Actionable Insights for Stakeholders: Design and build an administrator dashboard that provides a clear overview of citizen engagement, issue trends, and sentiment analysis. This objective ensures that the data collected is not just stored but is transformed into a valuable tool for data-driven decision-making, contributing to stronger, more transparent institutions.

### By achieving these objectives, the CivicVoice project will not only bridge the communication gap but will also serve as a model for how technology can be used to foster greater accountability and improve public welfare in developing communities.

### **Background**

### The Republic of Liberia, like many developing nations, faces a significant challenge in building resilient and accountable institutions. Decades of civil unrest and economic instability have left a legacy of underdeveloped infrastructure and a strained relationship between the government and its citizens. While the country has made strides in rebuilding, effective governance remains hindered by inefficient communication channels. Citizens often rely on informal networks or manual reporting systems that are slow, opaque, and offer no clear feedback loop. This leads to a public perception of inaction, a loss of trust, and a lack of data for policymakers to make informed decisions.

### **Existing Solutions and the Need for an AI Approach**

### Existing initiatives aimed at civic engagement in Liberia are often limited in scope or scale. Traditional solutions, such as community meetings and suggestion boxes, are valuable but are not scalable, often fail to capture diverse voices, and lack the data needed to track trends over time. More recently, some NGOs and government agencies have used social media platforms for outreach, but these platforms are unstructured, difficult to analyze for actionable insights, and can be sources of misinformation.

### 

### 

### A machine learning approach is not just beneficial—it is **necessary** to overcome these limitations. The sheer volume of unstructured, user-generated data (text, voice, and images) is too large and complex for manual analysis. A human team would be overwhelmed trying to read every complaint, categorize it, and identify emerging patterns.

### By using machine learning and Natural Language Processing (NLP), the CivicVoice platform can:

### **Process data at scale:** It can instantly analyze thousands of reports to identify keywords, topics, and sentiments in real time.

### **Generate actionable insights:** Instead of a pile of complaints, policymakers receive an analytics dashboard showing that, for example, "reports about sanitation in Monrovia increased by 30% this quarter," along with a heat map of the affected areas.

### **Enable a proactive response:** Predictive analytics can forecast future civic issues based on current trends, allowing authorities to intervene before a small problem becomes a larger crisis.

### In essence, a machine learning approach transforms the CivicVoice platform from a simple reporting tool into a powerful, data-driven system for improving governance and accountability. It provides a solution that is both efficient and scalable, a crucial requirement for addressing the challenges faced by a nation like Liberia.

### 

### 

### **Project Background**

### The country has been working to rebuild its infrastructure and institutions following years of civil war, but significant hurdles remain. Systemic challenges hinder effective governance and public service delivery. The CivicVoice AI app is a platform for citizens in Liberia to report urgent incidents and local issues. Using voice or text, people can submit reports for emergencies like crime, fire, and medical crises, as well as non-emergency problems like poor roads and power outages.

### 

### 

### 

### 

### 

### 

### 

### The key functions of NLP mentioned are:

### ● Sentiment Analysis: Automatically gauging the urgency and emotional tone of reports to flag critical situations.

### ● Keyword Extraction: Identifying important phrases like "armed robbery" or "road collapsed" to quickly get the core information from a report.

### ● Automatic Routing: Using the extracted keywords to automatically send a report to the appropriate department, like the police or fire service, to speed up response times.

### 

**Methodology**

The CivicVoice project will employ **machine learning and natural language processing (NLP)** techniques to process citizen feedback, classify issues, and generate actionable insights for policymakers and NGOs. The methodology includes:

1. **Data Preprocessing**

o Cleaning and normalizing text data (tokenization, stop-word removal, stemming/lemmatization).

o Handling multilingual input (Liberian English, indigenous languages).

o Converting voice submissions into text via speech recognition.

2. **Natural Language Processing (NLP)**

o **Sentiment Analysis:** Classifying feedback into positive, neutral, or negative categories using models such as Logistic Regression, Naïve Bayes, or Transformer-based models (BERT, RoBERTa).

o **Topic Classification:** Automatically categorizing reports into domains (education, healthcare, infrastructure, corruption, etc.) using Latent Dirichlet Allocation (LDA) or fine-tuned deep learning models.

o **Named Entity Recognition (NER):** Extracting key names, places, and organizations from citizen reports.

3. **Speech-to-Text Processing**

o Employing deep learning–based Automatic Speech Recognition (ASR) frameworks (e.g., Mozilla DeepSpeech, Wav2Vec2.0).

o Fine-tuning for Liberian English accents and dialects to improve accuracy.

4. **Geospatial Data Analysis**

o Using device GPS metadata for geotagging reports.

o Applying clustering algorithms (e.g., DBSCAN, K-means) to detect issue hotspots.

5. **Predictive Analytics**

o Using Gradient Boosting Machines (XGBoost/LightGBM) or Random Forests to forecast future civic issues.

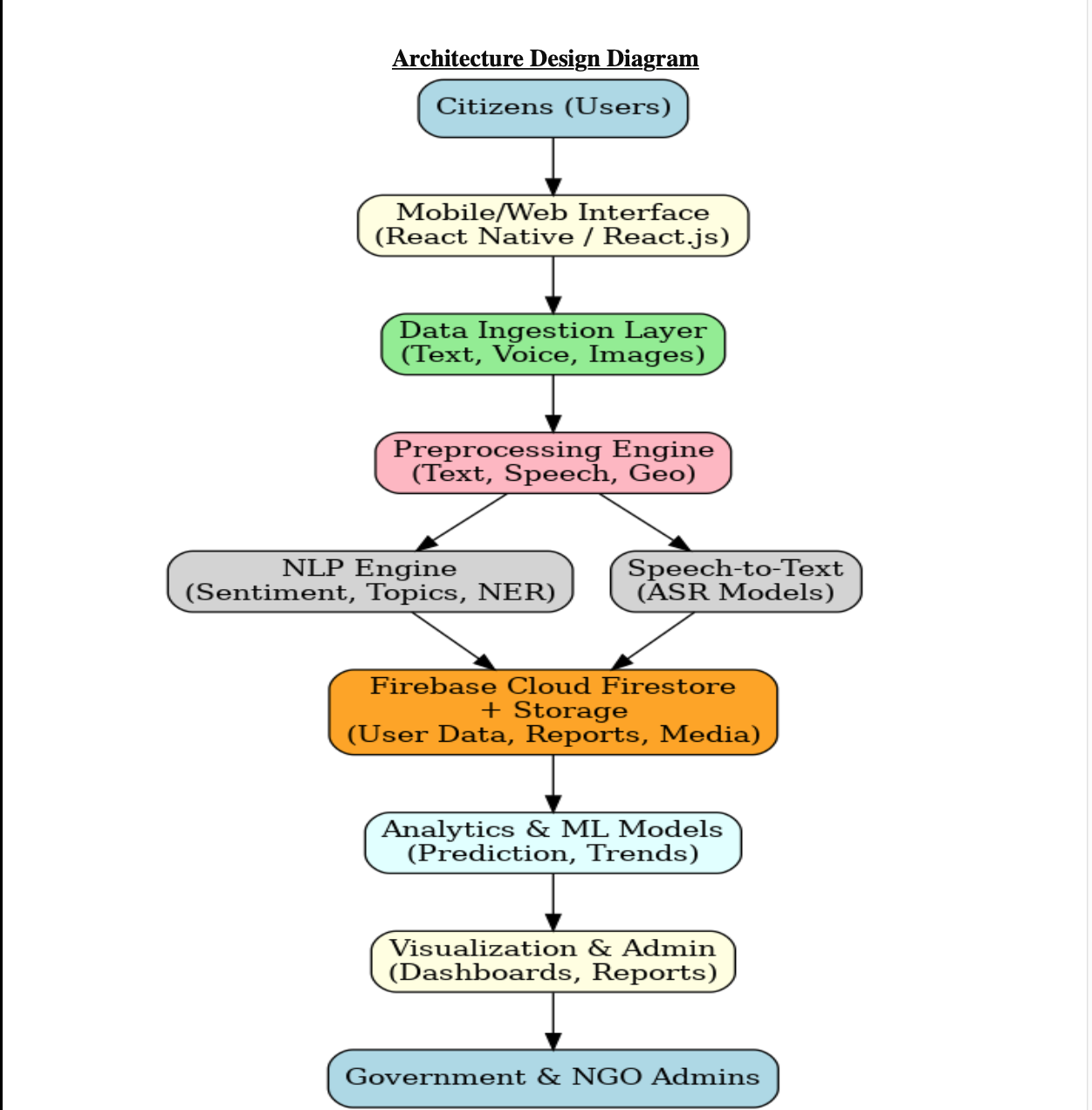
o Time-series models (ARIMA, Prophet, LSTMs) to identify trends in citizen engagement.

6. **Evaluation Metrics**

o Accuracy, Precision, Recall, and F1-score for classification.

o Word Error Rate (WER) for speech recognition.

o Engagement growth as a practical adoption metric.



**Architecture Design Diagram**

**Component Roles**

• **Citizens (Users):** Submit feedback via text, voice, or images.

• **Mobile/Web Interface:** Collects reports and provides user-friendly interaction.

• **Data Ingestion Layer:** Manages incoming text, voice, and geo-data.

• **Preprocessing Engine:** Cleans, transcribes, and attaches geotags.

• **NLP Engine:** Performs sentiment analysis, topic classification, and entity recognition.

• **Speech-to-Text Module:** Converts audio into text for analysis.

• **Firebase Database (Cloud Firestore):** Stores citizen reports, engagement logs, and user profiles. Firebase Storage handles media files (images, audio).

• **Analytics & ML Models:** Processes stored data to detect trends and generate predictions.

• **Visualization & Admin Panel:** Dashboards and reports for government and NGOs to act on insights.

**Data Sources**

The CivicVoice project will draw on both **primary** and **secondary** data sources. Primary data will consist of user-generated content collected directly from the platform, including posts, comments, survey responses, and engagement logs (likes, shares, votes). This type of data is highly relevant because it reflects real-time citizen concerns and levels of participation. Secondary data will include publicly available government reports, NGO publications, and open datasets (e.g., World Bank, Kaggle civic engagement datasets) that provide contextual benchmarks for validating findings. Since much of the primary data is unstructured text, preprocessing steps such as text cleaning, tokenization, stop-word removal, and lemmatization will be applied before analysis. Structured logs will also be standardized into tabular formats (CSV/SQL) to ensure consistency for modeling and visualization. Together, these data sources provide a comprehensive foundation for analyzing citizen sentiment, issue prioritization, and engagement patterns.

**Literature Review**

The proliferation of mobile technology and artificial intelligence (AI) has opened new avenues for enhancing public services, particularly in regions grappling with infrastructural and institutional challenges. This research is important because it seeks to leverage these technologies to bridge the communication gap between citizens and government, thereby improving public service delivery and governance. A comprehensive review of existing literature is necessary to establish the theoretical foundations for the CivicVoice project, identify successful strategies, and ensure the proposed solution contributes meaningfully to the field. This review will synthesize existing research on the application of AI, specifically Natural Language Processing (NLP), in public safety and governance to contextualize the importance and novelty of the CivicVoice app.

Artificial Intelligence (AI) has arisen as a driving technology to revolutionize various sectors, including [finance](https://www.sciencedirect.com/topics/social-sciences/finance), education, healthcare, and transportation. The advanced algorithms and [computational models](https://www.sciencedirect.com/topics/computer-science/computational-modeling) of AI are capable of imitating cognitive functions of human beings like learning, reasoning, and problem-solving. The [interdisciplinary nature](https://www.sciencedirect.com/topics/computer-science/interdisciplinary-nature) of AI encompasses subfields like [machine learning](https://www.sciencedirect.com/topics/computer-science/machine-learning), [natural language processing](https://www.sciencedirect.com/topics/computer-science/natural-language-processing), computer vision, and robotics, contributing to its rapid evolution and widespread adoption. Understanding the present landscape of AI research and development is crucial for harnessing its benefits while addressing ethical, societal, and [technical challenges](https://www.sciencedirect.com/topics/computer-science/technical-challenge).

The growing research in AI witnessed breakthroughs in [deep learning](https://www.sciencedirect.com/topics/computer-science/deep-learning), a subset of [machine learning algorithms](https://www.sciencedirect.com/topics/computer-science/machine-learning-algorithm) inspired by the structure and function of the human brain. AI development revolutionised tasks such as image recognition, speech recognition, and [natural language understanding](https://www.sciencedirect.com/topics/social-sciences/natural-language-understanding) through [deep learning techniques](https://www.sciencedirect.com/topics/computer-science/deep-learning-technique) like convolutional [neural networks](https://www.sciencedirect.com/topics/social-sciences/neural-network) (CNNs) and [recurrent neural networks](https://www.sciencedirect.com/topics/computer-science/recurrent-neural-network) (RNNs). The availability of large amounts of datasets and [computational resources](https://www.sciencedirect.com/topics/computer-science/computational-resource) has enabled the development of [AI models](https://www.sciencedirect.com/topics/computer-science/artificial-intelligence-model) resulting in remarkable performance in real-world applications. (Kulal et al., 2024)

### **Implementation**

Technology Stack

### **Technology Stack**

The CivicVoice project will leverage a robust and scalable technology stack to meet its objectives. The chosen technologies are well-suited for a real-world application, combining powerful machine learning capabilities with a reliable and user-friendly web and mobile interface.

#### Programming Languages

* Python: This will be the primary language for all backend and machine learning tasks. It is essential for building the NLP and predictive analytics models due to its extensive ecosystem of data science libraries.
* JavaScript/TypeScript: These languages will be used for the frontend development, handling the web and mobile interfaces.

#### Frameworks & Libraries

* Machine Learning & NLP:
  + Scikit-learn, TensorFlow/PyTorch: For building, training, and evaluating machine learning models.
  + Hugging Face Transformers: Crucial for implementing advanced NLP models like BERT or RoBERTa for more accurate sentiment analysis and topic classification.
  + Mozilla DeepSpeech or Wav2Vec2.0: For the Speech-to-Text module.
* Geospatial Analysis:
  + GeoPandas: A Python library for working with geospatial data.
  + Firebase Geoqueries and Google Maps API: For implementing location-based features and visualizing issue hotspots on a map.
* Data Handling:
  + Pandas, NumPy: For data manipulation and numerical operations.
* Frontend:
  + React Native: For building the cross-platform mobile application for Android and iOS.
  + React.js: For the web-based administrator dashboard.
* Backend & APIs:
  + Firebase Functions: To handle serverless, event-driven backend logic.
  + FastAPI / Flask: For creating APIs that expose the machine learning models.

#### Database & Storage

* Firebase Cloud Firestore: The primary database for storing all structured and semi-structured data, including user reports, profiles, and engagement logs. Its real-time capabilities are essential for live updates.
* Firebase Storage: For handling media files, such as audio recordings and images submitted by users.
* Google BigQuery (optional): A cloud data warehouse for large-scale analytics and long-term data storage.

#### Infrastructure

* Firebase Hosting: To host the web application and the mobile app's backend services.
* Docker (optional): For containerizing the machine learning models to ensure they run consistently across different environments.
* GitHub Actions: For Continuous Integration and Continuous Deployment (CI/CD) to automate the build and deployment process.

#### Hardware

* Smartphones (Android/iOS): For the pilot testing of the mobile application.

**Breakdown of Phases (12 Weeks)**

1. **Weeks 1–3:** Data Collection & Preprocessing

o Gather user feedback, surveys, and external datasets.

o Clean and preprocess text (tokenization, stop-word removal, normalization).

2. **Weeks 4–6:** Model Development

o Build classification models (Naïve Bayes, Random Forest).

o Implement NLP pipelines for topic modeling (LDA, NMF).

3. **Weeks 7–9:** Training & Evaluation

o Train sentiment analysis models (VADER, LSTM).

o Evaluate accuracy, precision, recall, and F1-score.

4. **Weeks 10–12:** Deployment & Testing

o Integrate models into CivicVoice platform.

o Test dashboard visualization and feedback loop.

o Final adjustments and documentation.

### 

### 

### 

### 

### **Key Metrics and Milestone for the CivicVoice Project so far**

We haven’t started development yet but so far, these are the milestones we have achieved.

### Phase 1: Planning and Scoping

* Milestone 1.1: Project Plan Finalized
  + Description: The project's scope, objectives, and high-level strategy are clearly defined and documented.
  + Metric: All key documents, including the Literature Review, Data Research, and Technology Review, are complete and reviewed by the team.
* Milestone 1.2: Technical Architecture and Stack Finalized
  + Description: The specific technologies and services to be used for the project are chosen, and the architectural design is fully mapped out.
  + Metric: The Architecture Design Diagram and Technology Stack documents are complete and approved.

### Phase 2: User and Stakeholder Research

* Milestone 2.1: Initial User Persona and Needs Defined
  + Description: Basic research is conducted to understand the target users (citizens and government officials).
  + Metric: A set of user personas and a list of their primary needs are documented.

### Phase 3: Data Sourcing and Preparation

* Milestone 3.1: Data Sourcing and Acquisition Strategy
  + Description: A clear plan for acquiring both primary (user-generated) and secondary (public) data is established.
  + Metric: All necessary agreements or API access for secondary data are secured, and a plan for primary data collection is drafted.

### 

### 

### 

### 

### Phase 4: Foundational System Setup

* Milestone 4.1: Development Environment Setup
  + Description: The core development environment is configured, including code repositories, CI/CD pipelines, and local development setups for all team members.
  + Metric: team members can successfully clone the repository and run a "Hello, World!" version of the planned application.

### **Anticipated Challenges and Mitigation Strategies**

### Here is an outline of the potential challenges that may arise during the project, along with strategies for addressing them.

#### **Data Quality**

### Challenge: User-generated primary data, such as comments and issue descriptions, can be inconsistent, contain typos, be incomplete, or include irrelevant content. This "noisy" data can reduce the accuracy of any analytical models.

### Mitigation Strategies:

### Implement robust preprocessing pipelines: Ensure the text cleaning and tokenization steps are comprehensive and can handle common issues like slang, misspellings, and special characters. For example, using libraries that include spell-checking or a thesaurus can normalize terms.

### Utilize data validation at entry: While you want to keep the user experience smooth, you can use simple checks to encourage better quality data, such as suggesting keywords or prompting users for more detail if a post is too short.

### Create a community moderation system: Allow users to flag posts for inappropriate content or misinformation. This crowdsourcing approach can help maintain data integrity and quality over time.

#### 

#### 

#### 

#### **Model Performance**

### Challenge: The unstructured nature of the primary data can make it difficult for machine learning models to accurately predict citizen sentiment or prioritize issues. A model might fail to understand nuance, irony, or sarcasm in a text. Furthermore, the model's accuracy might degrade as new types of issues or language trends emerge.

### Mitigation Strategies:

### Start with simpler models: Begin with straightforward models like logistic regression or a basic Naive Bayes classifier for initial sentiment analysis. These models are easy to interpret and serve as a strong baseline.

### Iterate with more advanced techniques: As the project grows, you can experiment with more complex models like transformer-based architectures (e.g., BERT) which are better at understanding the context of language.

### Implement a feedback loop: Allow users to rate the helpfulness of a model's output (e.g., "Was this issue tagged correctly?"). This human feedback can be used to continuously retrain and improve the model's performance.

#### **Technical Constraints**

### Challenge: The project requires handling and processing large volumes of data in real-time. Without a scalable architecture, the platform could become slow, unresponsive, or even crash under heavy user load.

### Mitigation Strategies:

### Leverage cloud-native services: Utilize services from cloud providers (like Google Cloud's BigQuery or AWS Redshift) that are designed for handling large datasets and can scale automatically based on demand.

### Optimize data storage: Instead of storing all raw data in a single database, use different storage solutions for different types of data. For example, a NoSQL database like Firebase is great for real-time user posts, while a structured warehouse is better for long-term analysis.

### Prioritize a simple, mobile-first design: This reduces the amount of data transferred and processed on both the client and server side, improving the overall speed and responsiveness of the application.

### 

### 

### **Ethical Considerations**

The Civic Voice project, by its nature as a platform for public discourse and data collection, must address several key ethical considerations to ensure it is responsible and beneficial to the community.

#### **Data Privacy**

Since the project handles user-generated content, protecting user data is paramount. The platform should be designed with a strong privacy policy that is easy for users to understand. This includes being transparent about what data is collected, how it is used, and who can access it. Measures to protect user identity are critical, especially for sensitive reports. For example, the system should allow users to post anonymously, and any publicly displayed data (like locations of reported issues) should be generalized or de-anonymized to prevent a specific individual from being identified. The project must also comply with data protection regulations such as GDPR or CCPA.

#### 

#### **Bias**

Bias can creep into the project at multiple stages. First, the data itself may be biased. The users who choose to participate in Civic Voice may not be representative of the entire community, potentially over-representing tech-savvy or more vocal groups. This could lead to a skewed understanding of which issues are most important to the community. To mitigate this, the project should actively seek to onboard a diverse user base and, when analyzing data, be transparent about the demographics of its users.

Furthermore, if the project uses machine learning models to analyze text or prioritize issues, there's a risk of algorithmic bias. The models could learn and amplify existing societal biases present in the training data, for example, by giving less weight to reports from certain neighborhoods or demographics. Regular audits of the models and the data they are trained on are essential to identify and correct any such biases.

#### 

#### 

#### 

#### 

#### **Impact on the Community**

The project's ultimate goal is to empower citizens, but it also has the potential for unintended negative impacts. For example, publicizing issues without a clear path to resolution could lead to frustration and distrust in local authorities. There is also a risk that the platform could be used to spread misinformation or be weaponized for targeted harassment against individuals or groups. The project must have clear moderation policies and community guidelines to prevent misuse. The design of the platform should encourage constructive dialogue and feedback, rather than simply being a complaint forum.

The project's success and ethical standing depend on a proactive and transparent approach to these issues, with a constant focus on building a platform that truly serves the public good.