Cohort 8 Group Members and Roles

1. Rachael Kibicho - Lead Data Science
2. George Karanja- Machine Learning Engineer
3. Angela Kinoro - Data Visualization and Reporting Specialist

Sidney Ochieng

Citizen Lens

Problem Background

Lack of transparency and accountability in government operations is widespread across Kenya. According to the Corruption Perceptions Index 2023 by Transparency International, Kenya scored 31 out of 100, ranking 123rd out of 180 countries, highlighting significant challenges in government transparency and corruption. This level of corruption in Kenya undermines Sustainable Development Goal 16, which targets building effective, accountable, and inclusive institutions at all levels. This accountability is expected in the allocation and management of funding disbursed by the local government, resources that are relied upon by all 47 counties in Kenya (ngcdf.go.ke). One of these sources of government funding is the Constituency Development Fund (CDF), established to address the unreliable progress of development initiatives and promote development at the constituency level and within local communities.

However, only 50% of NG-CDF projects are completed, 30% are implemented poorly and 20% still need to be completed at all funds are allocated. Furthermore, ghost projects and numerous public complaints about stalled or substandard projects are prevalent. (Anita et al., 2023, 2911). These challenges have a significant effect on residents of constituencies in Nairobi’s informal settlements, who rely on these development projects for better living conditions and essential services. Mathare, an informal settlement that faces additional challenges on CDF budget allocation received 37% less than the budgeted Kshs.214,505,337, totaling Kshs.135,498,218. Additionally, only 51% of these received funds were spent, resulting in a 49% under-expenditure. (2019/2020 Constituency Development Fund Audit Reports – Office of the Auditor-General, n.d.) This financial shortfall likely negatively impacted service delivery to residents. CDF projects address critical issues in these local communities, and the impact of failed or stalled projects is most acutely felt in these high-density populated areas.

The primary issue affecting the successful implementation of CDF projects is the lack of visibility. (Anandi Mani & Sharun W. Mukand, 2000) argue that when project outcomes are less visible, it becomes hard to access the government's effectiveness which may lead to underperformance and incomplete projects. While the National Government Constituencies Development Fund (NGCDF) website provides access to data about CDF projects, this information is often not regularly updated and lacks insights into the project progress of ongoing projects. Social media platforms like X, Instagram, and Facebook have also become popular tools for citizens to voice their concerns and share information about public projects, including CDF projects. However, these platforms are not designed to track CDF project progress or provide detailed accountability as these issues can often be drowned out by other content. This information is important because when citizens have access to relevant data, they can better understand government actions and increase scrutiny and pressure for these projects to be completed (Jenny, de, Fine, Licht. (2022)

How might we empower citizens interested in providing feedback about CDF projects, so that they keep their leaders and representatives accountable? How might we aggregate the needs of the constituents so that representatives ensure that CDF projects are completed and meet the actual needs of the constituents?

We aim to address this problem with Citizen Lens, a platform that enhances transparency and accountability in CDF projects through visibility and community engagement. Citizen Lens provides a USSD system where citizens may give feedback on the progress of any CDF project they wish. Furthermore, Citizen Lens will have an online system for the sentiments of citizens to be displayed to constituency representatives and the public after having passed through a sentiment analysis model. This will not only enhance the ability of citizens to audit their representatives' performance and project implementation efforts, but also enhance effective governance, and help voters make better-informed decisions during elections.

Market Opportunity

X, Facebook, and Instagram provide a space for citizens to express their views and concerns about government projects, including those funded by CDF funds. Social media platforms serve as a powerful platform for citizens to express their dissatisfaction with government practices.(Igbashangev et al. ,2023). In constituencies, citizens may rely on public forums and town hall meetings to express their dissatisfaction and concerns about the state of CDF projects.

While social media allows citizens to voice their concerns, there is no central platform specifically focused on allowing citizens to provide feedback on CDF projects. Social media posts tend to be fragmented and updates on CDF projects may easily be lost in the noise of other content. Additionally, there is no structured way of collecting, aggregating, and analyzing the feedback to draw meaningful insights and sentiments. Furthermore, these social media platforms require internet access, which excludes underserved areas where internet penetration is low. This limits citizens in these areas in providing feedback about CDF projects and participating in holding their leaders accountable.

CitizenLens employs USSD technology, allowing citizens who do not have internet access or smartphones to engage in providing feedback on CDF projects, giving a voice to the entire population. Additionally, CitizenLens is a dedicated platform that captures all CDF project feedback in one central system, ensuring that data is accessible, actionable, focused, and structured for CDF project monitoring

The primary target audience is residents of constituencies who rely on CDF projects for development, particularly those in informal settlements and underserved constituencies in Nairobi. For this study, we will focus on the Mathare constituency because of the striking underperformance in its (National Government Constituency Development Fund (NGCDF)) projects.

While this is primarily a civic engagement tool, there are potential revenue streams through NGOs, grants, sponsorships, partnerships, and other stakeholders who are interested in promoting visibility, transparency, and accountability in government operations and development projects. By enhancing visibility in the allocation and management of CDF funds, Citizen Lens can contribute to the economy by ensuring better implementation of projects, improved infrastructure, and overall quality of life in targeted constituencies. This leads to improved transparency and accountability thus reducing corruption and more effective use of public funds

Solution Idea

Target User

The primary users of Citizen Lens will be residents of the Mathare constituency. This area is selected due to its reliance on CDF funds for essential services like water and sanitation, health, infrastructure, and education.

While other constituencies also face a lack of transparency and accountability regarding NGCDF funds and projects, we chose to focus on Mathare, a highly populated, underserved constituency where NGCDF project funds have been acutely underutilized and where the solution will have a more significant impact on residents' well-being.

Solution Prototype

Solution offering and technology

Citizen Lens is a platform that uses a sentiment analysis model to aggregate the impressions that NGCDF projects have made on constituents in Mathare.

PostgreSQL is our chosen database technology. It is scalable and can handle large volumes of data coming from the USSD inputs. It also natively supports JSON which is a form in which our data will be at some point in processing.

Flask is our chosen web technology. It allows us to scale our web interface easily, and also integrate seamlessly with Python libraries, and can interact with the sentiment analysis model. Flask also allows the building of RESTful APIs to expose our sentiment analysis model. Flask is also being used to create the USSD logic, for the same compatibility and integration advantages.

Python is our machine-learning tool. It has extensive libraries that we are using, including, NLTK, HuggingFace, Pandas, and Scikit-learn all of which facilitated the deployment of the model. Python also integrates smoothly with Flask and with PostgreSQL enabling smooth flow of data from database, model, and web interface.

DistilBert model is the pre-trained model used. It has an accuracy of 90% to 93% with sentiment analysis tasks and an approximated 0.9 F1 score. This makes it great for use in our data, which is likely to be skewed negatively. It is a lightweight and fast model, making it ideal for our system.

We need the analyses to be made as quickly as possible. It has been pre-trained on a very large amount of text, so as to understand context and nuances which will definitely be present in our data. It can also be fine-tuned to fit a particular domain, like project progress and citizen auditing.

The dashboard uses React and Tailwind CSS since they facilitate very fast change accommodation. It is important to display the sentiments as soon as they have been analyzed.

Africa’s Talking API is being used for USSD services because it is very scalable, secure through encryption, and allows for secure data storage. Africa’s Talking is also very well supported by developers, including integration, troubleshooting, and optimizing USSD applications.

Dummy data from Sci-kit Learn is being used to train, test, and validate due to time constraints and the short development time.

Solution Process

A USSD logic has been set up at the input, using Flask. The user enters the USSD CODE : *789*9085635# and is taken to the logic. The logic takes the user through a series of questions that they answer. The answers are stored in the PostgreSQL database in readiness for analysis by the model.

The data from the PostgreSQL is loaded. The data is tokenized by DistilBertTokenizerFast. The model is then set to evaluation mode. The input data is run through the model and the dictionary of inputs is unpacked and passed as keyword arguments to the model. The model returns its outputs, which are in logits. These logits are converted by the softmax function into probabilities that sum to 1. This results in a probability distribution over the probable classes. If the probability of class 1 (positive) is higher, then the model returns 1, and if the probability of class 0 (negative) is higher, then the model returns 0.

The sentiments of the users are presented in an aggregated format in a dashboard. The dashboard is designed to enhance transparency by displaying the derived sentiments through visualizations such as pie charts. This output aims to give citizens and constituency representatives a clear and impactful overview of the public sentiment of Mathare Constituency.

How the solution solves the problem DIRECTLY:

Citizen Lens addresses the visibility gap in CDF projects by providing a centralized, easy-to-use platform that empowers citizens to ensure the completion and accountability of CDF projects.

Assumptions made:

User1(Citizens)

1. The users understand and can type in English.

2. The users have a phone.

3. The users have access to a cellular network.

User2(Constituency representatives)

1. The users have access to the internet.

2. The users have computer literacy and have devices to access the internet.

Value proposition

For Mary, a dedicated mother in Mathare, our platform gives her the power to voice her concerns about local projects like hospital services. By amplifying her feedback, we ensure that authorities stay accountable, improving essential services that directly impact her family's well-being and the broader community.

Designed Solution

Technologies Used

Africa's Talking API: This API is used to power the USSD service that residents use to provide their feedback - reliability, and ease of integration with local telecom services.

Flask: building our REST API, which handles the USSD requests and sentiment analysis - simplicity and ease of integration with other technologies, allowing for quick development and deployment.

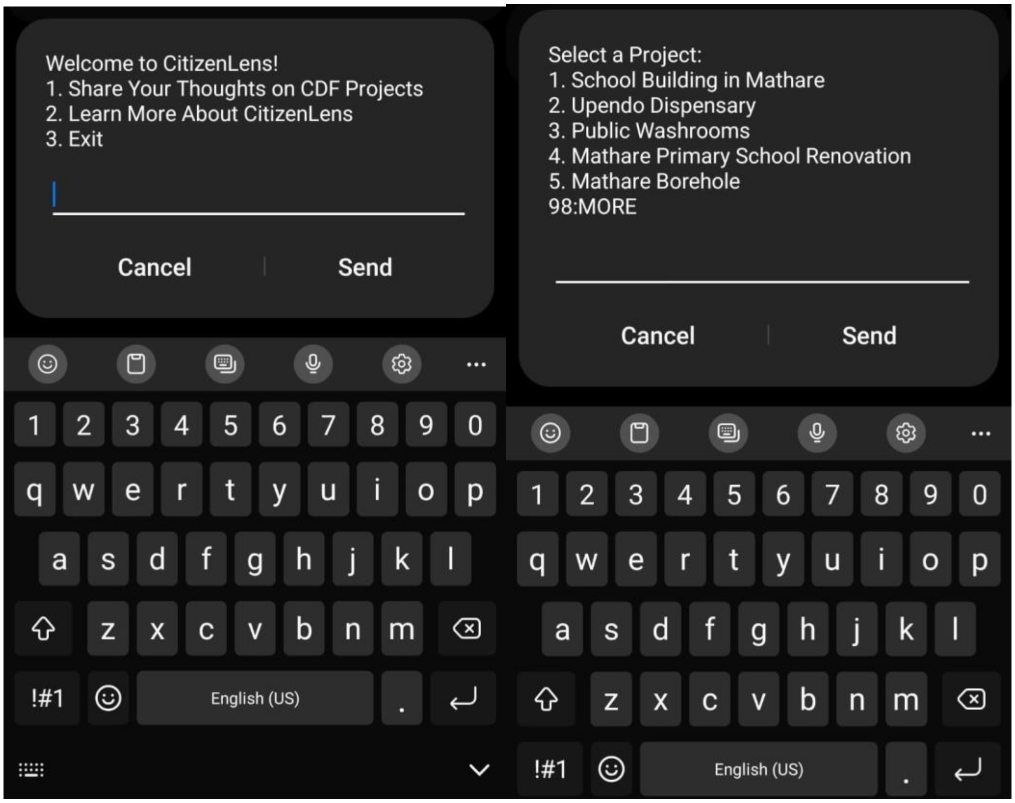
PostgreSQL: PostgreSQL is the database used to store all feedback from residents, the sentiment results, and project details. Its robustness, scalability, and ability to handle complex queries made it an ideal choice for securely storing and managing our data.

DistilBERT: We used a pre-trained **DistilBERT** model from Hugging Face for sentiment analysis. It provides state-of-the-art text classification with minimal computational resources, making it suitable for real-time analysis of feedback data.

React & Tailwind frameworks: These are the front-end technologies used to build a user-friendly dashboard that displays aggregated feedback results. It enables local authorities to easily access insights on how residents perceive different projects.

Main Modules

1. **USSD Feedback Module**

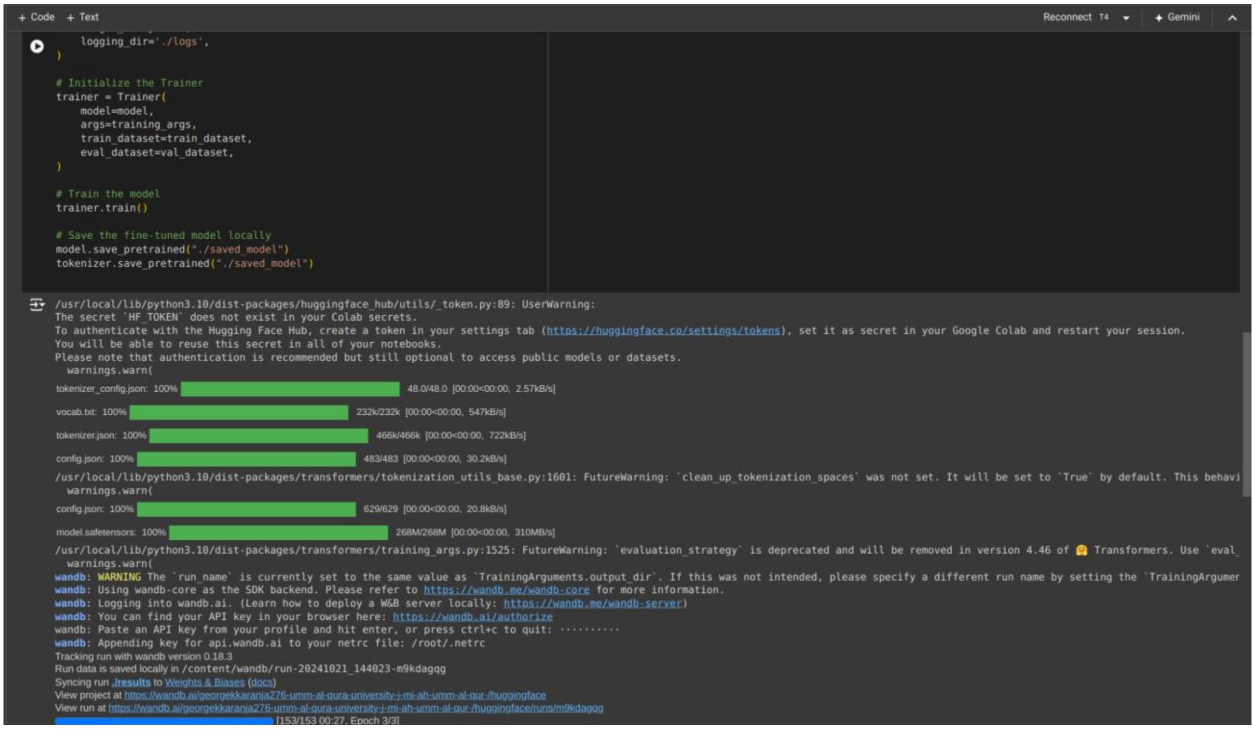


Welcome to CitizenLens!
1. Share Your Thoughts on CDF Projects
2. Learn More About CitizenLens
3. Exit

Select a Project:
1. School Building in Mathare
2. Upendo Dispensary
3. Public Washrooms
4. Mathare Primary School Renovation
5. Mathare Borehole
98:MORE

This module allows residents to provide feedback by selecting a community project and describing its current state and impact on their lives.

2. Sentiment Analysis

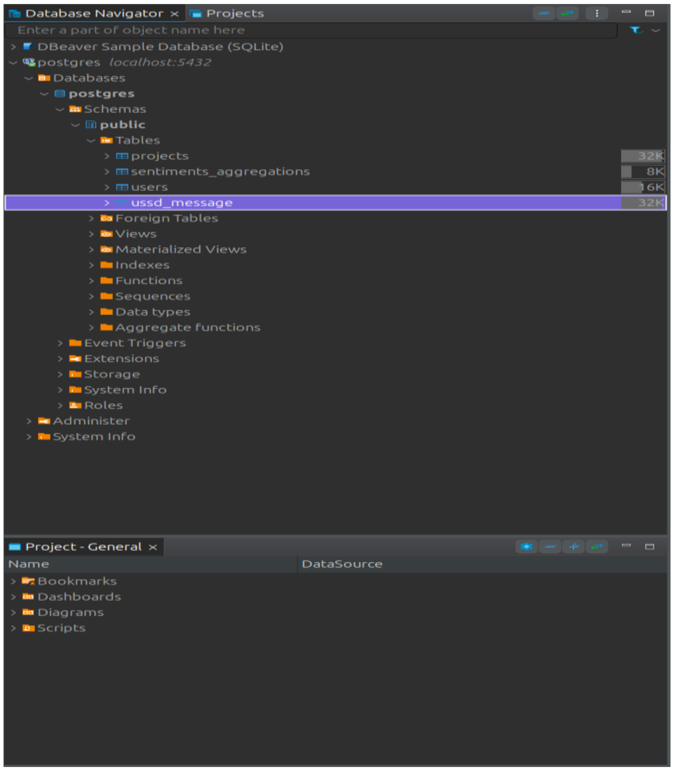


```
    logging_dir='./logs',
)

# Initialize the Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=val_dataset,
)

# Train the model
trainer.train()

# Save the fine-tuned model locally
model.save_pretrained("./saved_model")
tokenizer.save_pretrained("./saved_model")
```

```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
tokenizer_config.json: 100%  48.0/48.0 [00:00<00:00, 2.57kB/s]
vocab.txt: 100%  232k/232k [00:00<00:00, 547kB/s]
tokenizer.json: 100%  466k/466k [00:00<00:00, 722kB/s]
config.json: 100%  483/483 [00:00<00:00, 30.2kB/s]
/usr/local/lib/python3.10/dist-packages/transformers/tokenization_utils_base.py:1601: FutureWarning: `clean_up_tokenization_spaces` was not set. It will be set to `True` by default. This behavi
  warnings.warn(
config.json: 100%  629/629 [00:00<00:00, 20.8kB/s]
model.safetensors: 100%  268M/268M [00:00<00:00, 310MB/s]
/usr/local/lib/python3.10/dist-packages/transformers/training_args.py:1525: FutureWarning: `evaluation_strategy` is deprecated and will be removed in version 4.46 of 🤗 Transformers. Use `eval_
  warnings.warn(
wandb: WARNING The `run_name` is currently set to the same value as `TrainingArguments.output_dir`. If this was not intended, please specify a different run name by setting the `TrainingArgumer
wandb: Using wandb-core as the SDK backend. Please refer to https://wandb.me/wandb-core for more information.
wandb: Logging into wandb.ai. (Learn how to deploy a W&B server locally: https://wandb.me/wandb-server)
wandb: You can find your API key in your browser here: https://wandb.ai/authorize
wandb: Paste an API key from your profile and hit enter, or press ctrl+c to quit: ··········
wandb: Appending key for api.wandb.ai to your netrc file: /root/.netrc
Tracking run with wandb version 0.18.3
Run data is saved locally in /content/wandb/run-20241021_144023-m9kdagqg
Syncing run ./results to Weights & Biases (docs)
View project at https://wandb.ai/georgekkaranja276-umm-al-qura-university-j-mi-ah-umm-al-qur-/huggingface
View run at https://wandb.ai/georgekkaranja276-umm-al-qura-university-j-mi-ah-umm-al-qur-/huggingface/runs/m9kdagqg
 [153/153 00:27, Epoch 3/3]
```

After the feedback is received, it is passed through the sentiment analysis model, which classifies the text as positive or negative, helping to gauge public sentiment.

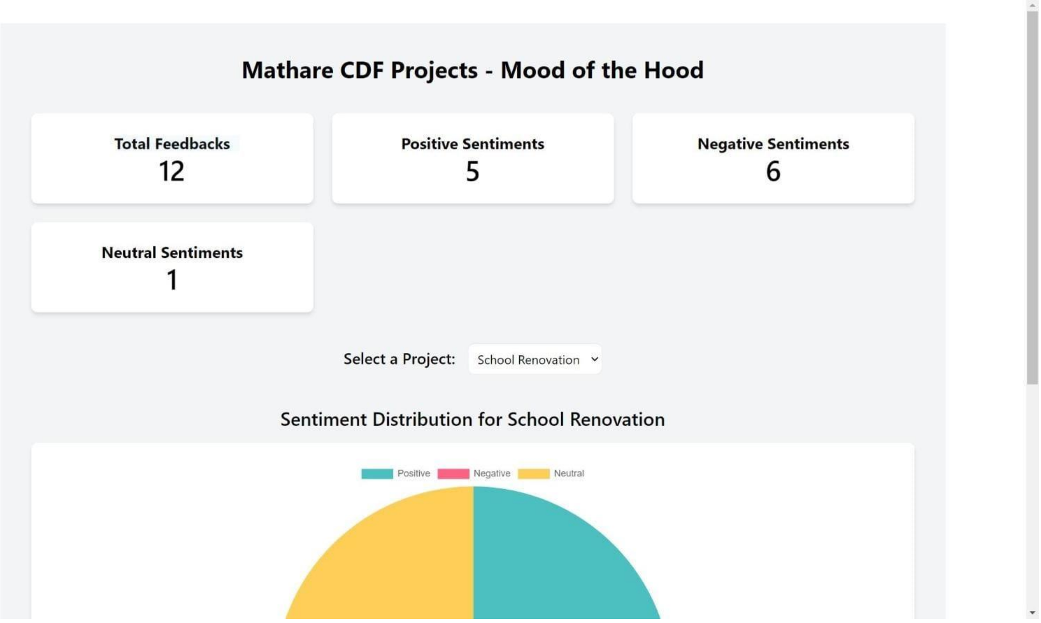
3. Database Storage



```
Database Navigator × | Projects
Enter a part of object name here
> DBeaver Sample Database (SQLite)
v postgres  localhost:5432
  v Databases
    v postgres
      v Schemas
        v public
          v Tables
            > projects                      32K
            > sentiments_aggregations        8K
            > users                         16K
            > ussd_message                  32K
          > Foreign Tables
          > Views
          > Materialized Views
          > Indexes
          > Functions
          > Sequences
          > Data types
          > Aggregate functions
        > Event Triggers
        > Extensions
        > Storage
        > System Info
        > Roles
  > Administer
  > System Info

Project - General ×
Name                                  DataSource
> Bookmarks
> Dashboards
> Diagrams
> Scripts
```

This shows how the feedback and sentiment results are stored in the database. We use relational mapping to connect the feedback to specific projects.

4. Dashboard for Aggregation



Mathare CDF Projects - Mood of the Hood

Total Feedbacks
12

Positive Sentiments
5

Negative Sentiments
6

Neutral Sentiments
1

Select a Project: School Renovation

Sentiment Distribution for School Renovation

This web interface displays aggregated feedback results for different community projects, giving authorities a clear picture of the public's sentiment.

Link to the solution

GitHub Link:

https://github.com/angelakinoro/KamiLimu-Sentiment_Analysis

Executable mode:
https://tense-bulb.surge.sh/

Business Model

As a non-profit initiative aimed at improving public service transparency and accountability, we intend to raise funds through strategic partnerships, sponsorships, and grants from organizations committed to fostering civic engagement and government transparency. Our primary focus will be on building relationships with large organizations, NGOs, and foundations that are aligned with our mission of empowering communities like Mathare. These partners can offer financial support and resources to help scale our platform.

Responsible Computing

Citizen Lens includes all constituents who have a feature phone. This is following the exclusion of this demographic in political discourse, due to a lack of resources to acquire devices that allow for usage of applications such as X, Facebook, and Instagram.
Citizen Lens is also able to give a voice to those individuals who have little, unstable, or no access to the internet This is because as recently as the beginning of 2022, Kenya's internet penetration rate stood at 42.0% of the total population (DataReportal's Digital 2022: Kenya report).

Citizen Lens uses open-source Kenyan tweet data from Kaggle to fine-tune the DistilBert model. The use of this data is intentional since we wanted the model to understand the context of Kenyan culture from the kind of language contained in the tweets. The use of this data to fine-tune the model also removes the bias contained in the Western original tweet data that was used to train the model initially.

Traction

Our solution is yet to be used by an individual from our target audience. It is at the post-pilot phase that we intend to have users interact and give feedback on the impact Citizen Lens will have on their experience of their constituency leadership and workmanship.

Funding/Support Need

Digital Marketing and Community Outreach (Ksh. 3,600,000 for 3 years)

This is based on an average digital marketing cost of 100,000 shillings and an average cost of Event Organization ranging from 100,000 to 300,000 shillings.

Supporting Software (Ksh.1,377,000 for 3 years)

Render database hosting

Year 1 & 2: approximately KES 36,000/year.

Year 3: approximately KES 90,000/year
To scale up hosting tiers as the user base grows

Heroku USSD Application Hosting

Year 1: approximately KES 45,000/year

Year 2: approximately KES 90,000/year

Year 3: approximately KES 180,000/year

Africa's Talking (Charges 1 KES per session)

Year 1: 100,000 sessions * KES 1 = KES 100,000

Year 2: 200,000 sessions * KES 1 = KES 200,000

Year 3: 400,000 sessions * KES 1 = KES 400,000

Increase of session usage every year.

Installation of USSD code with Safaricom: sh. 200,000

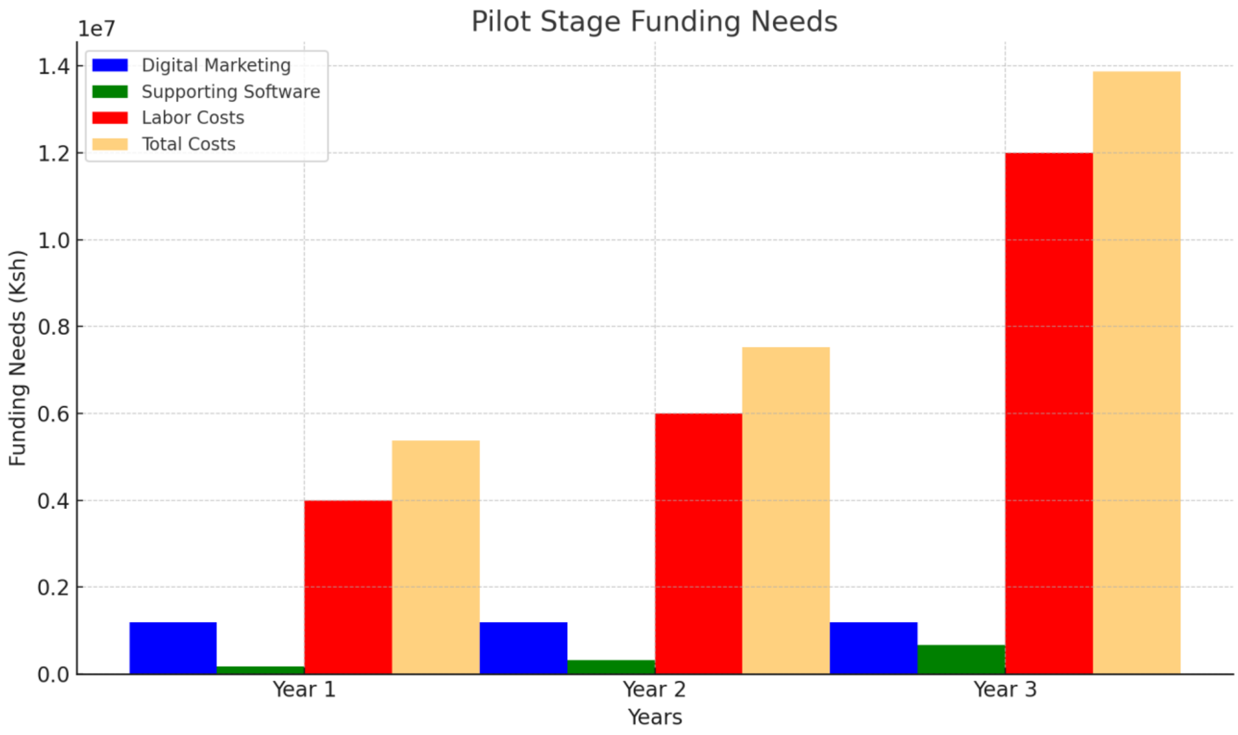
Labor costs(Ksh. 3,600,000 for 3 years)

Year 1: KES 4,000,000 for 2 data scientists, 1 web developer, and a project manager.

Year 2: KES 6,000,000 for 2 data scientists, 1 web developer, a project manager, and a board of decision-makers.

Year 3: KES 12,000,000 for 2 data scientists, 1 web developer, a project manager, a board of decision-makers, and support staff including human relations specialists.

Total: Ksh. 6,100,000 for Pilot phase



Your team

The Citizen Lens team is composed of two data scientists and one front-end web developer:

Rachael Kibicho

The team lead and lead data scientist. She has worked with various data science experts to tackle Africa's greatest problems using data science tools for instance at Zindi Competitions.

George Karanja

A vibrant machine learning engineer with experience at Women In Tech Huawei

Angela Kinoro

A phenomenal web developer currently dedicated to making great web interfaces for companies like Mama Pesa at Chandaria Innovation Hub.