Truth or Deception: A Multi-Class NLP Approach to Detecting Fake News in Kenya

Background & Motivation.

The media landscape in Kenya has significantly evolved from traditional print and broadcast to include a strong digital presence. With this evolution—particularly the rise of social media—Kenya has experienced a surge in both misinformation and disinformation across online platforms, especially during critical periods such as elections, national disasters, and political unrest.

According to the 2025 *Reuters Institute Digital Report (RIDR)*, 59% of Kenyans believe that online influencers and political figures are the leading source of news distortion in the country, identifying them as the biggest threat to news accuracy.

The ease of access to online publishing has also led many individuals to self-identify as journalists. Combined with limited media literacy among some users, this has created an environment where false narratives can rapidly circulate and influence public opinion.

The rapid spread of fake news not only erodes trust in the media but also poses serious risks to democratic processes, public safety, and social cohesion.

Despite global efforts to combat misinformation, there remains a shortage of locally contextualized tools that can automatically detect and classify fake news within the Kenyan media ecosystem. This project seeks to address that gap by leveraging Natural Language Processing (NLP) and Machine Learning to develop a multi-class fake news detection model tailored specifically for Kenyan news content.

By integrating real-time web scraping from both credible and questionable Kenyan news websites, this project offers a grounded and relevant solution to a challenge that affects millions of Kenyans every day. It not only sharpens our technical capabilities but also contributes to the growing demand for media literacy and truth-detection tools in East Africa.

Problem Statement

The rise of fake news in Kenya—particularly during politically sensitive periods—has resulted in widespread misinformation, public confusion, and a growing erosion of trust in mainstream media. At present, there is no publicly available machine learning model that can accurately detect and classify Kenyan news articles based on their truthfulness.

According to the latest *Reuters Institute Digital Report* (RIDR), based on online responses from English-speaking Kenyans aged 18–50, **73% of respondents** reported struggling to differentiate true from false information online. This concern is particularly acute across Africa, where misinformation often spreads rapidly during key national events.

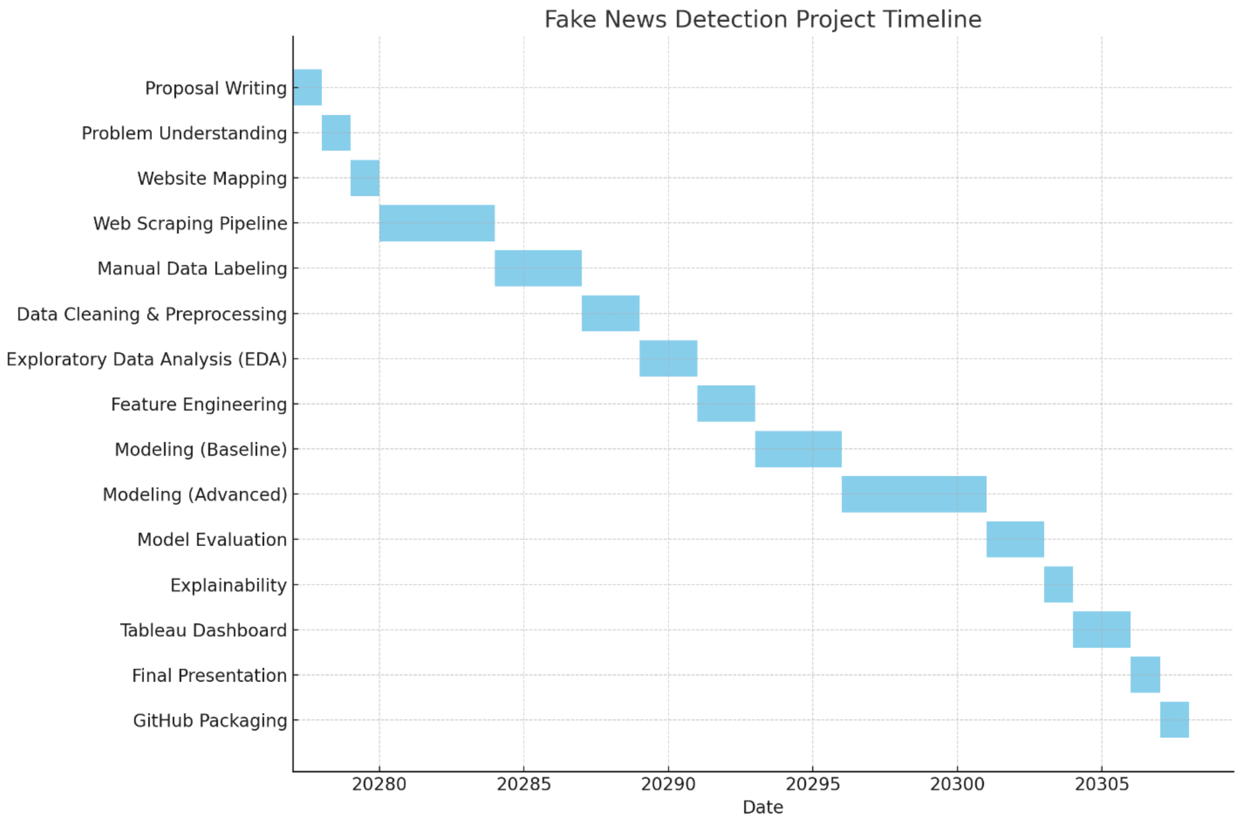
While fake news detection has been widely studied at the global level, most existing models fail to account for the **cultural, linguistic, and contextual nuances** specific to Kenya. This project aims to bridge that gap by building a localized, multi-class classification model capable of identifying not only true and fake news, but also **satire, clickbait, and misleading content** within Kenyan digital news platforms

Project Objectives

1. **Develop** a multi-class classification model to categorize Kenyan news articles into: *True, Fake, Satire, Misleading*.
2. **Collect & label** a sizable dataset with at least:
    - 2,000 examples per category from diverse Kenyan news websites.
    - Include authentic fake-news cases (e.g. Ndiangui, Mwangi deepfake).
3. **Compare** traditional techniques (e.g., Logistic Regression, SVM) vs. deep learning (LSTM, BERT).
4. **Implement** explainability tools (SHAP/LIME) to interpret model predictions.
5. **Deploy** results via:
    - A **Tableau dashboard** visualizing class breakdown, site credibility, trends.
    - A **non-technical presentation** for stakeholders.
    - A **GitHub repo** with README, requirements.txt, usage guide.
6. **Evaluate** model performance based on precision, recall, and F1 per class, aiming for **≥ 70% macro-F1**

Milestone	Description	Estimated Duration
1. Proposal Writing	Define project title, background, objectives, scope, methodology	1 day
2. Problem Understanding	Research fake news categories in Kenya, define final class labels	1 day
3. Website Mapping	Identify 6–10 Kenyan news websites (credible + questionable) for scraping	1 day
4. Web Scraping Pipeline	Build and test scraping scripts (Selenium/Requests + BeautifulSoup)	3–4 days
5. Manual Data Labeling	Clean raw articles and manually label a seed set (~500–1000 articles)	2–3 days
6. Data Cleaning & Preprocessing	Text normalization, handling missing values, tokenization	1–2 days
7. Exploratory Data Analysis (EDA)	Word clouds, term frequencies, length distributions, comparison plots	1–2 days
8. Feature Engineering	TF-IDF, n-grams, stylistic features, entity count, etc.	2 days
9. Modeling (Baseline)	Train traditional ML models (LogReg, NB, SVM) for quick benchmark	2–3 days
10. Modeling (Advanced)	Train deep learning models (LSTM, BERT), fine-tune & compare	3–5 days
11. Model Evaluation	Analyze confusion matrix, per-class F1, visualize results	1–2 days
12. Explainability	Integrate SHAP or LIME to interpret key predictions	1 day
13. Tableau Dashboard	Design an interactive visualization of the findings	1–2 days
14. Final Presentation	Prepare a non-technical summary with story, visuals & impact	1 day

Milestone	Description	Estimated Duration
15. GitHub Packaging	Polish README, push code/data, generate requirements.txt	1 day



Scope

- The project focuses on **classifying Kenyan news articles** into multiple truthfulness categories: *True, Fake, Satire, Misleading*, and possibly *Clickbait*.
- Only **textual content** will be analyzed — no images, videos, or audio.
- News articles will be scraped from a **curated list of Kenyan news websites**, both reputable and questionable, as well as social news pages if feasible.
- The solution will include both **traditional machine learning** and **deep learning NLP models**.
- A sample-based **manual labeling strategy** will be used to bootstrap the dataset.

Limitations

- Due to the **absence of labeled public datasets** specific to Kenyan news, the dataset will be **partially self-labeled**, which may introduce bias or inconsistency.
- Not all fake news websites openly label their content, making it difficult to guarantee full accuracy in label assignment.
- The model may struggle with **code-switching** (e.g., Swahili-English) or **sarcasm/satire**, especially in short or informal articles.
- Deep learning models (e.g., BERT) may be limited by **computational resources** unless run on cloud platforms or Google Colab Pro.
- The project will not address **real-time detection** or **news propagation patterns** on social media — only the classification of standalone articles.

Methodology

This section outlines exactly how we'll approach the fake news classification problem, step by step:

Step 1: Data Collection

- Identify and scrape news articles from multiple Kenyan news websites.
  - *Credible sources:* Nation, Standard Media, Citizen Digital.
  - *Suspected or satirical sources:* Blogs, lesser-known political news sites.
- Extract metadata: title, date, full text, source URL.
- Store all data in a structured format (e.g., SQLite or CSV).

Step 2: Data Labeling

- Manually label a seed dataset (500–1000 articles) across 4–5 categories:
  - *True, Fake, Satire, Misleading, Clickbait* (optional).
- Use heuristics or weak supervision (e.g., keyword patterns or source reputations) to expand labeling semi-automatically.

Step 3: Text Preprocessing

- Clean article text:
  - Remove HTML tags, punctuation, stopwords, and lowercasing.
  - Tokenize and lemmatize (using spaCy or NLTK).
- Handle class imbalance (e.g., SMOTE, class weights).
- Optional: Detect and translate Swahili snippets for consistency.

Step 4: Exploratory Data Analysis (EDA)

- Visualize:
  - Word frequency per class (True vs Fake).
  - Article length distributions.
  - Source reliability heatmaps.
  - Topic trends across categories.

Step 5: Feature Engineering

- Use both classical and linguistic features:
  - TF-IDF vectors, n-grams, text length, named entity counts.
  - Source trust score (based on manual review).
  - Readability scores (Flesch-Kincaid).
- Optionally include title sentiment or capital letter frequency for satire/clickbait detection.

Step 6: Modeling

- **Baseline Models:** Logistic Regression, Naive Bayes, SVM.
- **Deep Learning:** LSTM (with embedding), BERT (fine-tuned for text classification).
- Split data into training, validation, and test sets (e.g., 70–15–15).
- Train and compare models using macro F1-score, precision, recall, and confusion matrices.

Step 7: Model Explainability

- Use SHAP or LIME to interpret why the model classifies an article a certain way.
- Show which words or phrases most influence predictions.

Step 8: Dashboard and Communication

- Build an interactive **Tableau dashboard** to:
    - Show class distributions by source.
    - Highlight trends in fake/satire content.
    - Include SHAP/LIME visual summaries.
- Prepare a **non-technical presentation** explaining how the tool works, why it matters, and sample outputs.

Step 9: Packaging

- Document the project thoroughly:
    - Clean README.md with problem, methodology, results, and instructions.
    - Push source code, datasets (or scraping scripts), visuals to GitHub.
    - Generate requirements.txt for reproducibility.

Expected Outcomes

By the end of this project, we expect to achieve the following:

---

1. A Fully Functional Fake News Classifier

- A trained, validated multi-class classification model capable of accurately categorizing Kenyan news articles as:
  - **True**
  - **Fake**
  - **Satire**
  - **Misleading**
  - *(Optional: Clickbait)*
- Macro F1-score target: **≥ 70%** on test set.

---

2. A Real-World, Locally Relevant Dataset

- A labeled dataset of at least **2,000–3,000** articles from Kenyan news sources.
- Collected using web scraping, covering multiple media outlets and styles.

---

3. Transparent, Interpretable Predictions

- SHAP/LIME visualizations that explain **why** a news article is classified as fake or satire.
- Highlights most influential words or patterns.

---

4. An Interactive Tableau Dashboard

- Visual analysis of:
  - Distribution of fake vs. true content across sites.
  - Topic or keyword clusters per class.
  - Timeline trends and deep-dive article stats.

---

5. A Complete Communication Package

- **Non-Technical Presentation** for stakeholders explaining:
  - Why fake news matters
  - What our model does
  - Real examples from Kenyan news
- **GitHub Repository** with:
  - Project code & scripts
  - Clean README.md
  - requirements.txt for easy setup
  - Sample predictions and analysis

---

6. Improved Mastery of NLP & Model Deployment

- Hands-on experience with:
  - Multi-class NLP modeling
  - Feature engineering for fake news detection
  - Explainability tools
  - Project documentation best practices
- New project to showcase on LinkedIn, GitHub, and your portfolio