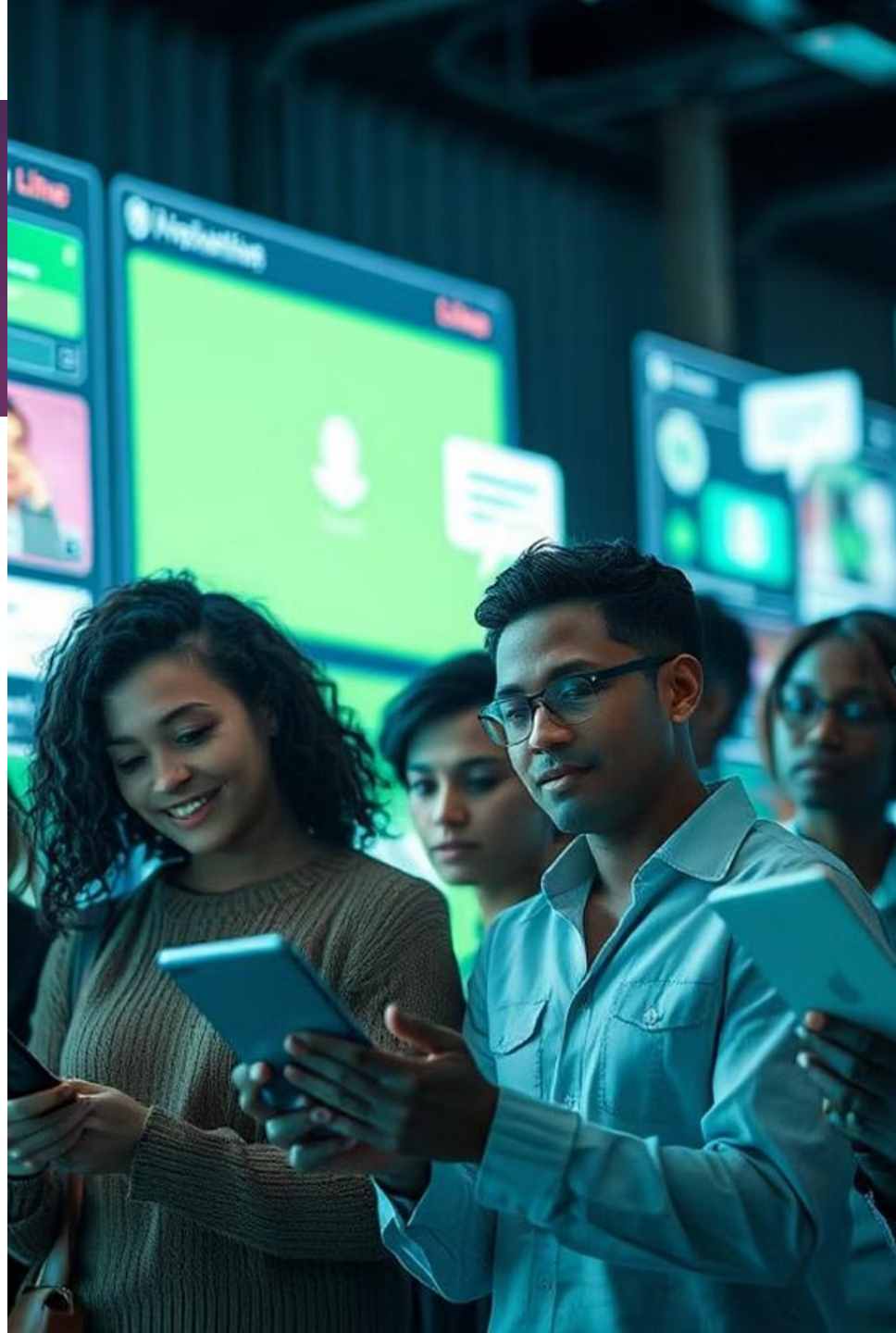
Detecting Hate Speech in the Digital Age

- An ML-driven approach to identifying and mitigating harmful content.
- Team: Vincent Ngochoch, Chris Gitonga, Kelvin Kibet, Shawn J. Irungu



Why This Matters Now

- • Protect vulnerable groups from harassment and discrimination.
- • Prevent online speech from escalating to real-world violence.
- • Maintain platform trust, safety, and reputation.

Kenya's Unique Challenge

- • Twitter is a key platform for political discourse in Kenya.
- • Risks: Hate speech, tribalism, and incitement during elections.
- • Challenge: Manual moderation is slow, biased, and misses local nuances.
- • Solution: Automated ML classification model for early detection.

Our Mission & Goals

- • Identify common inciting terms and phrases.
- • Build a multi-class classification model (Hate Speech, Offensive, Neutral).
- • Evaluate algorithms using F1 Score.
- • Deploy best-performing model for real-time detection.

Inside the Data

- • Kenyan election-related hate speech dataset from Kaggle.
- • Labels: Hate Speech, Offensive Language, Neutral.
- • Thousands of annotated tweets for robust training.

How We Clean the Data

- • Remove irrelevant characters, URLs, tags, and mentions.
- • Tokenization into words/sub-words.
- • Stop word removal and lemmatization.
- • Standardizing text for better model understanding.

Turning Text into Features

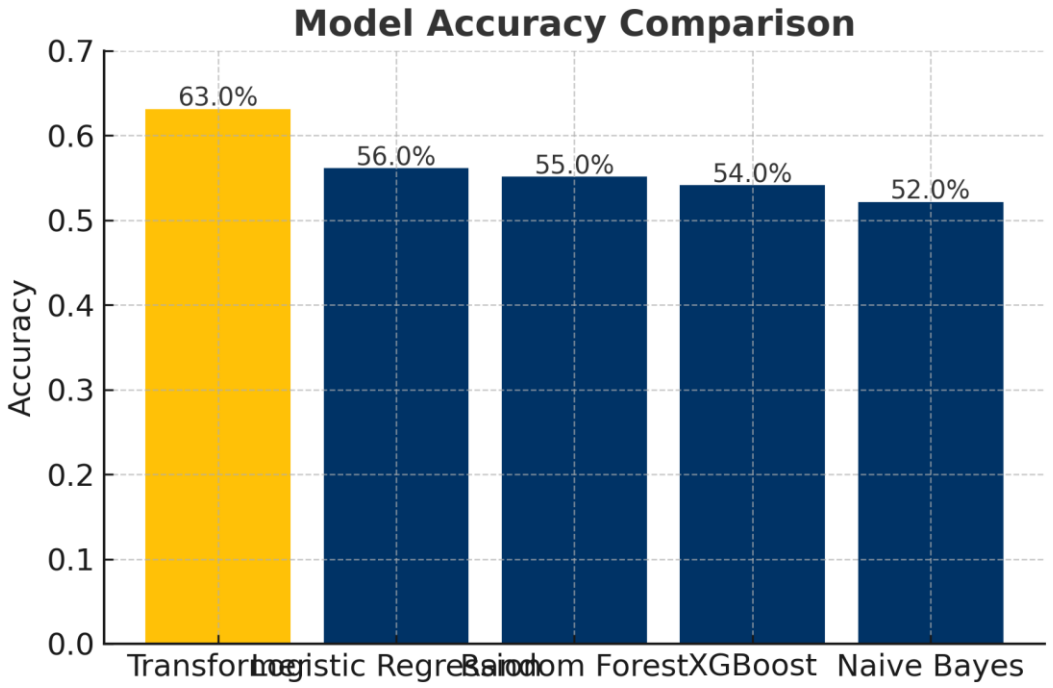
- • TF-IDF scores: Highlight important terms.
- • Sentiment polarity: Detect emotional tone.
- • Word embeddings: Capture semantic meaning.
- • Lexical features: Text length, emoji presence.

Models We Tested

- • Logistic Regression – strong baseline, efficient.
- • Naive Bayes – probabilistic text classifier.
- • Random Forest – rule-based tree model.
- • XGBoost – high-performing ensemble method.
- • RoBERTa Transformer – contextual deep learning model.

Which Model Wins?

- • Logistic Regression: Best overall accuracy.
- • RoBERTa: Strong on Hate Speech & Neutral detection.
- • Both models show strong balance between precision and recall.



From Model to Action

- • Deploy best model via Streamlit web app.
- • Real-time tweet classification.
- • User-friendly dashboard for moderators.

What We Learned

- • ML models can detect harmful speech with high accuracy.
- • TF-IDF and sentiment analysis are valuable features.
- • Scalable system suitable for high-volume platforms.

Where We Go From Here

- • Expand dataset for broader coverage.
- • Improve model for detecting nuanced, local expressions.
- • Integrate into social media monitoring systems.