# SWAHILI NEWS CLASSIFICATION MODEL.

# Project overview

- **Contracting Party**: A news airing company based in East Africa.
- **Team Involved**: Group Six, consisting of five members.
- **Objective**: Increase the company’s sales potential.
- **Focus Area**: Expansion within the news category to be aired.
- **Task Scope**: Enhance content and strategies to boost revenue and audience reach.
- **Location**: East Africa, a region with growing media demand.
- **Goal**: Strengthen the company’s market position through improved sales and influence.

# Business understanding

- **Challenge in Media Reporting**: Managing and organizing large volumes of Swahili news content.
- **Manual Classification Issues**: Time-consuming and prone to inconsistencies.
- **Need for Automation**: Crucial for news platforms, media houses, and content aggregators

# Data understanding

- **Dataset Columns**:
- **id**: Unique identifier for each article.
- **content**: Swahili text of the article.
- **category**: News article category (e.g., Uchumi, Kitaifa, Michezo).
- **Categories**:
- **Uchumi**: Business, finance, and economic news.
- **Kitaifa**: General news about Tanzania.
- **Michezo**: Sports-related news.
- *(Other categories may exist).*

# Problem statement

•**News Media Role**: Shapes opinion, informs societies, and influences discourse.
•**Swahili in East Africa**: Unifying language for news in Tanzania, Kenya, Uganda,
•Rwanda, Burundi, and DRC.
•**Digital Journalism Growth**: Rising Swahili news content needs automated classification.
•**Project Goal**: Use NLP and Deep Learning to categorize Swahili news.

# Objectives

● **Automating News Classification.**

How can we efficiently categorize Swahili news articles using machine learning and deep learning?

● **Understanding Media Trends**.

What are the most common news topics in East African media?

Are there any biases in media coverage based on classification trends?

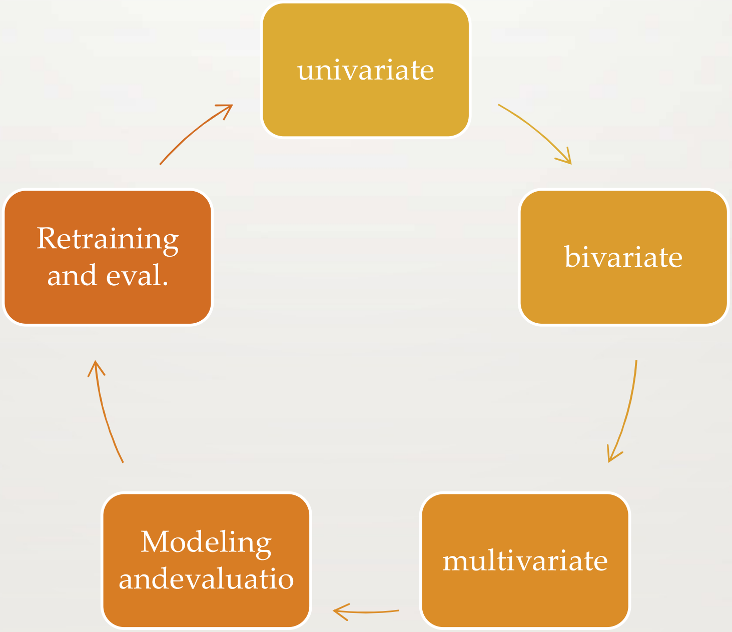
● **Enhancing Content Accessibility.**

How can automated classification improve information retrieval for journalists, policymakers, and the general public?

# methodology

- Rule based methods –POS
- Statistical methods-Naïve bayes
- ML models-SVM, Random forests
- Deep learning models-RNN,BERT
- Vector space models-TF-IDF

# workflow



univariate

bivariate

multivariate

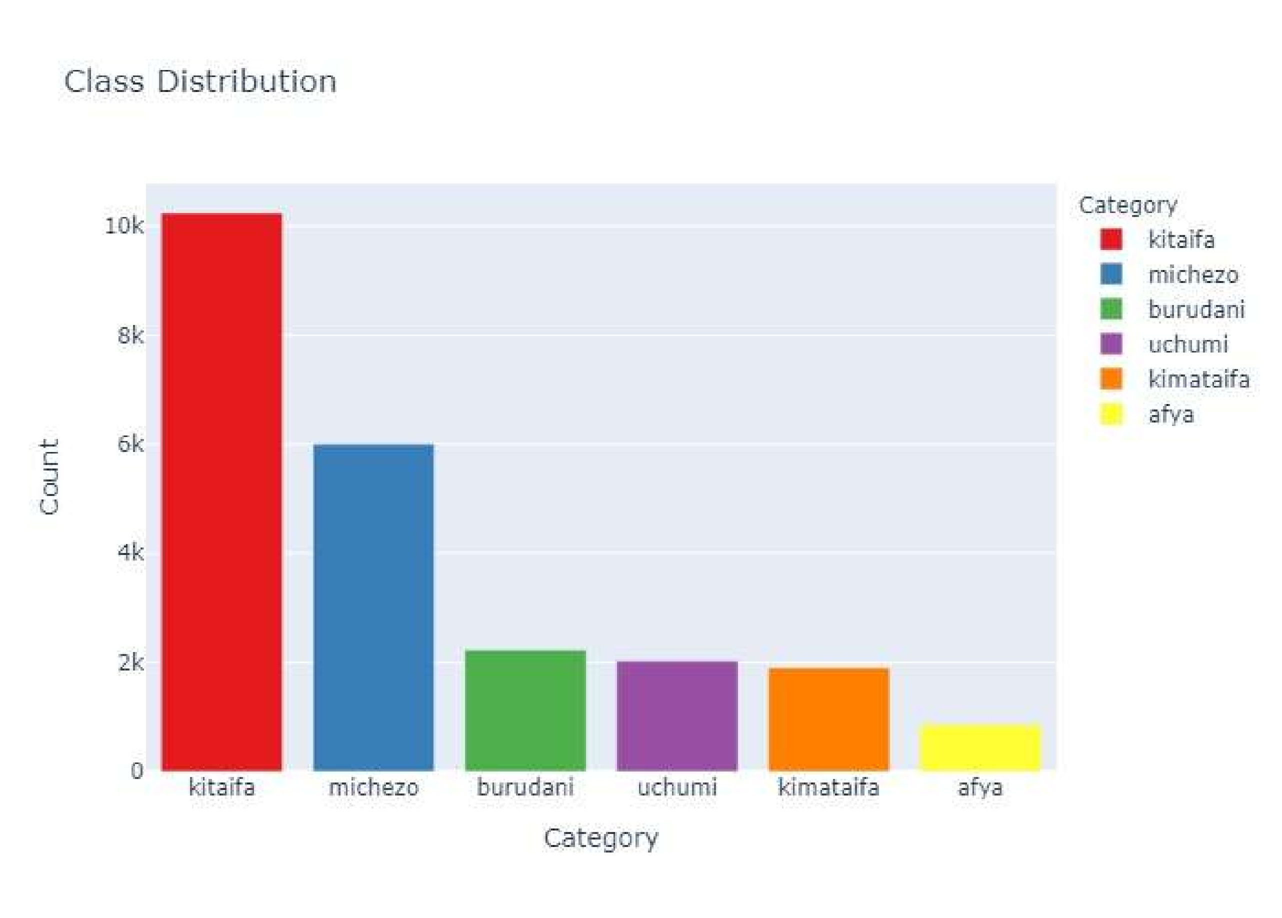
Modeling andevaluatio

Retraining and eval.

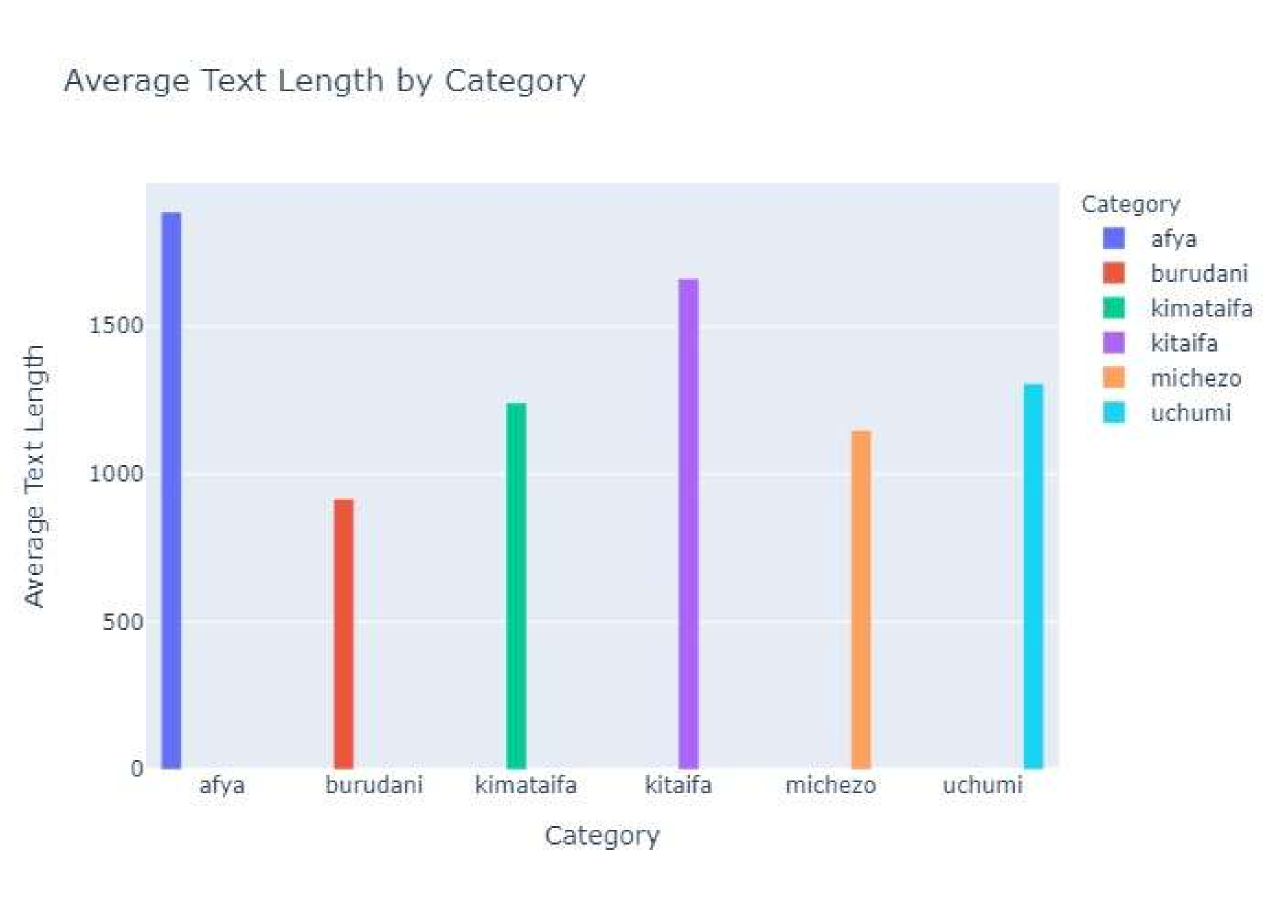
# Exploratory data analysis

- Univariate
- Bivariate
- Multivariate

# Class distribution for the various categories



# Average text length by category



# Modeling report and evaluation (Fine-Tuned BERT-base-uncased for Sequence Classification)

1. High Accuracy (89.41%)

2. High F1 Score (89.34%)

3. The close alignment between accuracy (89.41%) and F1 score (89.34%) indicates that the model is consistent in both positive and negative predictions.

4. The strong validation performance and high F1 score suggest that the model generalizes well to unseen data, not just memorizing the training set.

# Conclusion

- The "afya" and "kitaifa" categories have the longest average text lengths.
- The "afya" category has the highest average word count, indicating longer and possibly more complex content.
- Best classifier is Fine-Tuned BERT-base-uncased for Sequence Classification with 89.34% F1 score.

# Recommendations

- Focus on creating balanced content. Shorter and more concise for categories like "burudani" and more detailed for categories like "afya" and "kitaifa."

- Tailor content length based on user preferences to improve retention and interaction.

- Automate content tagging and classification to improve searchability and content recommendations using the **Fine-Tuned BERT-base-uncased for Sequence Classification**

# Acknowledgement

- Technical mentor: Mr.William Okomba
- Team members: group 6 members

# Questions

- THANK YOU!