Swahili News Classification Project

GROUP 7

- John Nkakuyia
- Miriam Ongare
- George Tido
- Mercy Ronoh
- Shalom Irungu

Business Problem

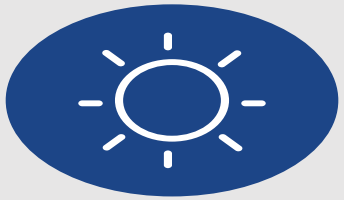
In Tanzania and across East Africa, Swahili serves as a vital language for communication, education, and cultural expression. However, with the increasing dominance of English in online spaces, there's a risk of losing the representation of Swahili, especially in digital media such as news platforms. This project aims to address this challenge by developing a multi-class classification model to automatically categorize Swahili news articles into specific categories. By doing so, online news platforms can enhance user experience by providing readers with easy access to news content relevant to their interests, while also contributing to the preservation and promotion of the Swahili language in the digital age.

Objectives

Questions to Consider

1

Develop a Multi-Class Classification Model: Utilize machine learning techniques to accurately classify Swahili news articles into predefined categories, including Kitaifa, Kimataifa, Biashara, Michezo, and Burudani.



2

Enhance User Experience: Improve the accessibility of Swahili news content by enabling automated categorization on online news platforms, facilitating easier navigation for readers.

3

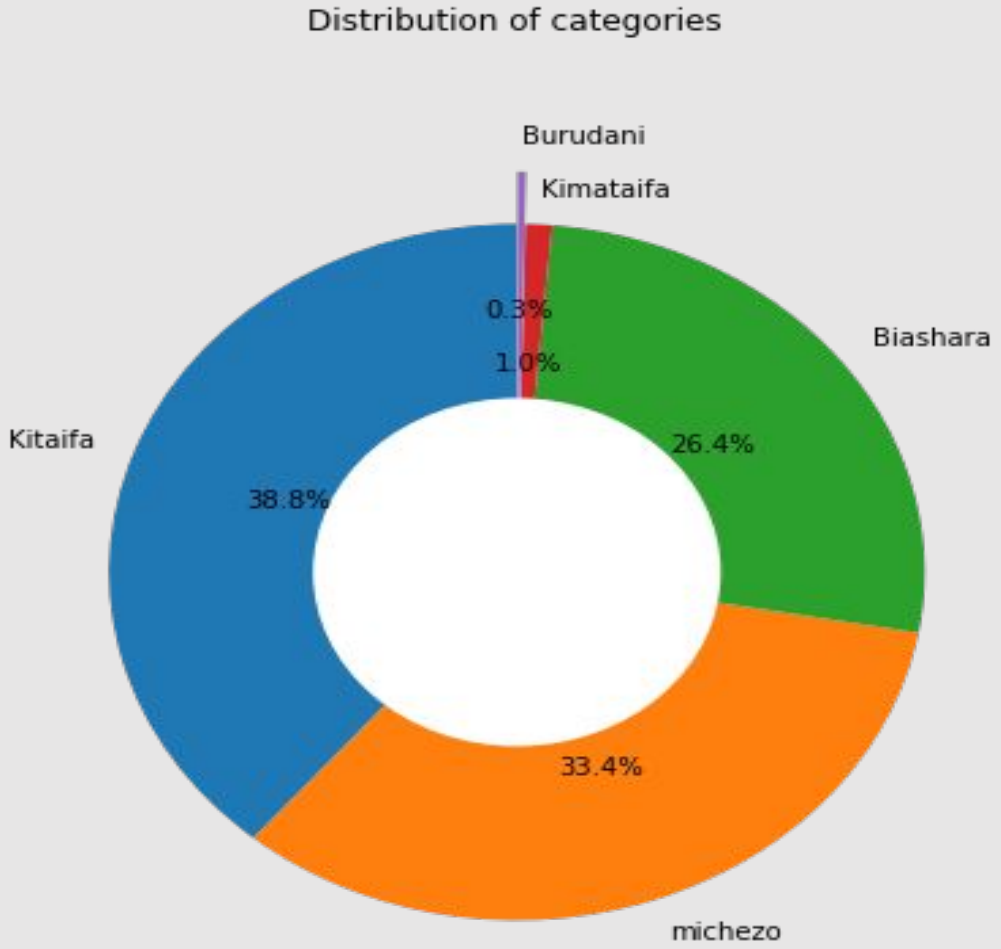
Promote Swahili Language: Contribute to the representation and preservation of Swahili in digital media by ensuring its inclusion and visibility in online products and services.

Explore Data

- Data Source: Zindi Afrika- news article
- Features: id, content, category
- Distributed across 5151 rows and 3 columns
- Target variable: Category
- Category 'Kitaifa' , 'Biashara', 'michezo', 'Kimataifa' , 'Burudani'

DATA PREPROCESSING & EDA:

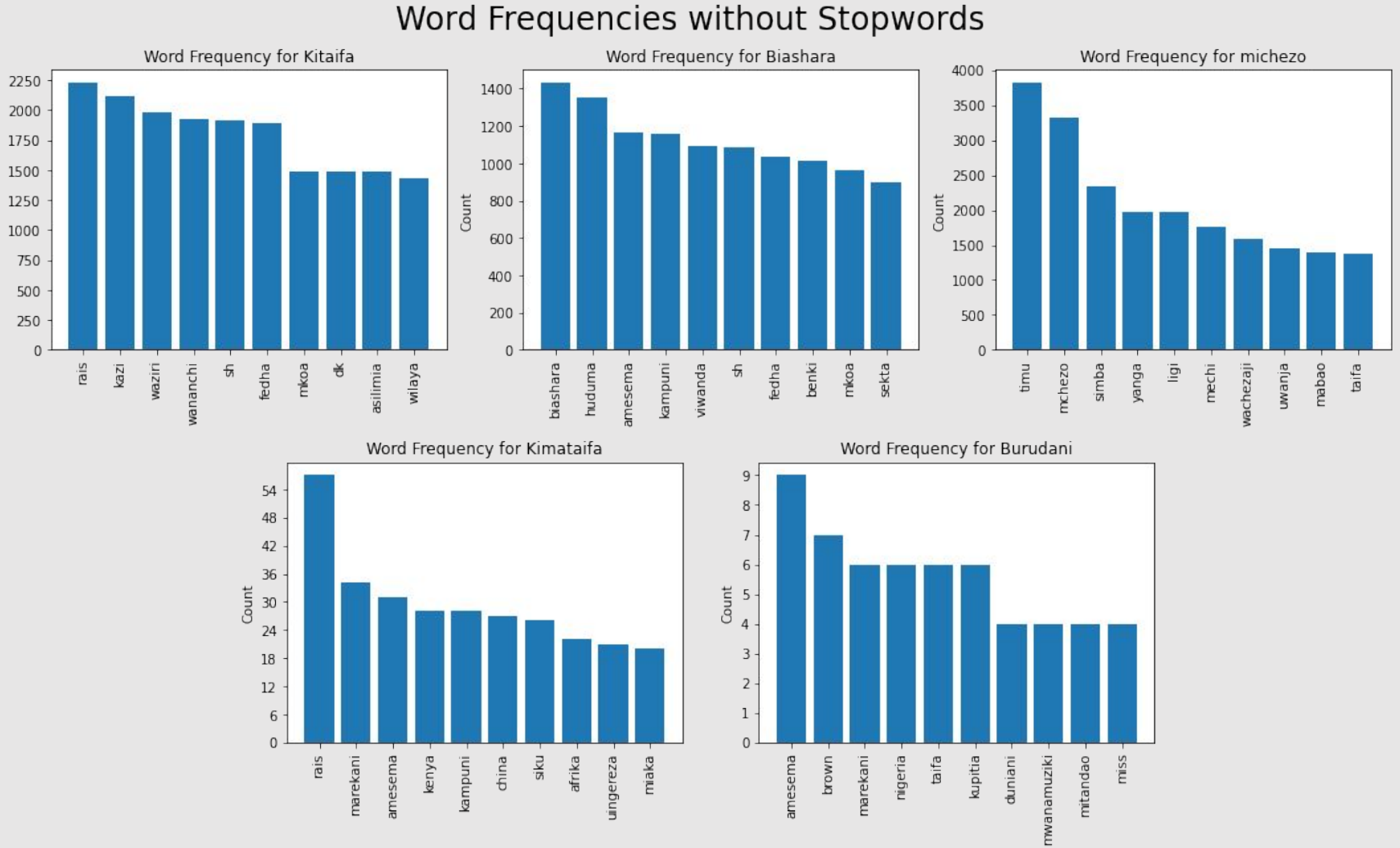
Category Distribution



The top 3 categories in the news articles are:

- Kitaifa : 38.8%
- Mchezo : 33.4 %
- Biashara : 26.4%

Frequency Distribution for the 5 news categories

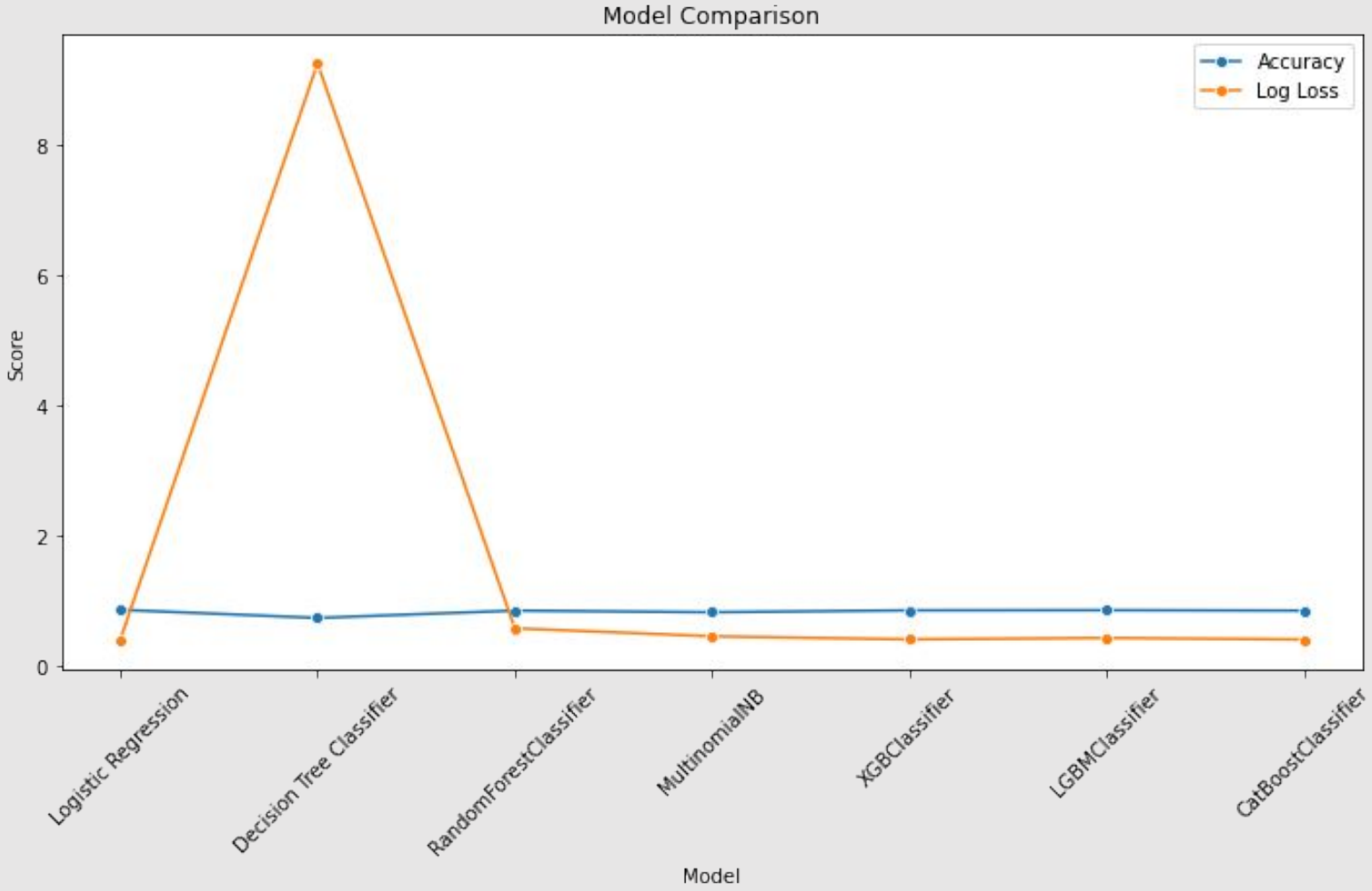


Frequent words with no stopwords:

- rais
- fedha
- mkoa
- taifa
- kampuni
- marekani

MODELLING

Model comparison



- LGBMClassifier achieved the highest accuracy of 0.880,
- CatBoostClassifier with an accuracy of 0.861.
- LGBMClassifier having slightly higher precision and recall.
- CatBoostClassifier shows slightly higher accuracy.

Evaluation

- Rare classes like Burudani and Kimataifa pose challenges for the model due to their limited presence in the dataset, leading to lower true positive values in the confusion matrix.
- Stacking model achieved 85.24% accuracy, indicating that the majority of its predictions were correct, and a log loss of 0.3984, suggesting close alignment with actual probabilities.
- Analysis shows consistent and precise predictions across all classes without mislabeled posts, indicating the model's reliability for accurate classifications.
- The model's performance indicates its potential suitability for real-world deployment where accurate classification is vital.

Conclusion

- Rare classes pose challenges for accurate classification due to limited data representation.
- Stacking model demonstrates strong performance with 85.24% accuracy and minimal misclassifications, indicating reliability for real-world deployment.
- Further analysis is needed to ensure model robustness across diverse datasets and identify any biases or limitations.

Recommendations

- Conduct additional testing with varied datasets to assess model generalizability.
- Implement bias detection mechanisms to identify and mitigate potential biases in the model.
- Continuously monitor model performance in real-world applications to ensure sustained accuracy and reliability.

Swahili News Classification Project

THANKS FOR WATCHING!