ANALYSIS OF EAST AFRICAN SWAHILI NEWS

MORINGA DS -FT 11

GROUP VI PROJECT

PRESENTED BY :

1. Rodgers Ndemo
2. Catherine Kipkirui
3. Amani Amkaya
4. Adnan Mohamud
5. Bernice Kutwa

TABLE OF CONTENTS

ANALYSIS OF EAST AFRICAN SWAHILI NEWS

1. PROJECT OVERVIEW

News media plays a crucial role in shaping public opinion, informing societies, and influencing political and social discourse. In East Africa, Swahili is one of the most widely spoken languages, serving as a unifying medium for news dissemination across multiple countries, including Tanzania, Kenya, Uganda, Rwanda, Burundi, and the Democratic Republic of Congo.

With the rise of digital journalism, there has been a rapid increase in Swahili news content, necessitating the need for automated classification of Swahili news articles. This project aims to leverage Natural Language Processing (NLP) and Deep Learning to develop a model that can accurately categorize Swahili news content into predefined categories

1.2.1 Business overview

The challenge in media reporting is managing and organizing large volumes of Swahili-language news content efficiently. Manual classification is time-consuming and prone to inconsistencies, making automated categorization crucial for news platforms, media houses, and content aggregator

1.2.2 Problem statement

The goal of this project is to build a Swahili news classification model that accurately categorizes news articles into six predefined categories:

- uchumi (economy)
- kitaifa (national news)
- michezo (sports)
- kimataifa (international news)
- burudani (entertainment)
- afya (health)

To achieve this, we will preprocess the text data to remove noise, tokenize, and normalize the text, followed by building a classification model. To ensure a clear, reproducible, and scalable approach, we will implement the preprocessing steps within an Scikit-learn Pipeline.

1.2.3 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?

2. DATA UNDERSTANDING

2.1 Overview

This dataset consists of Swahili news articles categorized into different topics. The goal is to classify news articles into predefined categories using Natural Language Processing (NLP) and Deep Learning techniques.

2.1.0 Data Structure

The dataset contains the following columns:

id: A unique identifier for each news article.

content: The text of the news article written in Swahili.

category: The label representing the category of the news article (e.g., uchumi, kitaifa, michezo).

2.2.1 Categories in the Data

The dataset has multiple categories representing different types of news. Some of the common categories include:

Uchumi (Economy): Articles related to business, finance, and economic activities.

Kitaifa (National News): General news related to Tanzania.

Michezo (Sports): News about sports teams, events, and athletes.

(Other categories may exist and need to be explored further.)

2.2.2 Data Size and Distribution

To understand the dataset better, key aspects to analyze include:

The total number of articles.

The distribution of articles across different categories (class imbalance analysis).

The average length of articles in terms of word count.

2.2.3 Data Quality Issues

Potential issues to check before preprocessing:

Missing values: Are there any missing or empty fields?

Duplicates: Are there repeated news articles?

Class imbalance: Are some categories significantly overrepresented compared to others?

Noise in text: Presence of irrelevant characters, symbols, or stopwords that may need cleaning.

2.2 Data Collection

2.3 Data Summary

2.4 .Exploratory Data Analysis

2.5 verifying Data Validity

3. DATA PREPARATION

3.1 Data cleaning

4. DATA ANALYSIS

5. RECOMMENDATIONS

6. CONCLUSION