WELCOME TO THIS COURSE

NLP & African Language Models -DSA 2026

ABIGAIL OPPONG | LOISE MWARANGU

6 Modules | 58 Slides | Self-Paced Learning

By the end of this course, you will understand NLP fundamentals, build pipelines, and apply transfer learning to African languages.

Overview

- The Foundation: Translating messy human language (text and audio) into computer logic.
- Core Applications: Training AI for Sentiment Analysis, Summarization, NER, and Voice Assistants.
- The Pipeline: Data Ingestion → Cleaning (Pre-processing) → Vectorization (Math) → Model Training.
- The Toolstack: Using Python libraries (spaCy, NLTK, Scikit-Learn, Librosa) powered by Google Colab.
- Responsible AI: Fighting bias with inclusive local datasets and practicing Green AI.
- Transfer Learning & BERT: Reusing massive, pre-trained HuggingFace models to drastically save time and reduce data needs.

MODULE 1

Introduction to NLP

Learning Objectives:

- Understand what Natural Language Processing is
- Explore why NLP matters for African languages
- See real-world examples of NLP in action

What is NLP?

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that helps computers understand, interpret, and generate human language.

Computer Reality
Computers naturally only understand math **(1s and 0s)**.

The NLP Bridge
It translates English, Swahili, Hausa, or Zulu into a language the computer can process.

NLP in Africa: Bridging the Language Gap

Africa has thousands of languages and dialects. While standard AI models are built for Western countries, NLP allows us to build tools that understand local slang like **Sheng** or **Pidgin**.

The Messy Challenge

Human language breaks rules constantly, creating hurdles for computers:

Ambiguity:
"I hit the bat with a bat." (Animal vs. Tool)

Context:
"The bank is closed." (River vs. Finance)

NLP Solutions

Building technology that actually understands us:

- **Weather via SMS:** Helping local farmers access critical data.
- **Localized Bots:** Customer service that speaks your dialect.
- **Cultural Nuance:** Decoding regional slang and idioms.

The Challenge: Sarcasm & Slang

Humans use emotion and sarcasm that a dictionary cannot explain.

"If a matatu breaks down and you tweet: 'Oh great, exactly what I wanted today!'"

Standard Computer: Reads "great" and "wanted" and thinks you are happy.
With NLP: Helps the computer realize you are actually annoyed by understanding the sarcastic context.

Everyday NLP Applications

Where do we see this technology working today?

Finance

Extracting transaction details (amount, date, recipient) from SMS messages for budgeting.

Chatbots

Instant customer support on WhatsApp, understanding user intent to provide correct information.

Spam Filtering

Recognizing suspicious patterns and language to classify emails as spam.

KNOWLEDGE CHECK

Module 1: Introduction to NLP

1. What does NLP stand for, and what is its main purpose?

2. Give two examples of how NLP can help African communities.

3. Why is sarcasm difficult for computers to understand?

4. Name three everyday applications of NLP mentioned in this module.

Take a moment to reflect on these questions before moving on.

MODULE 2

Core Applications

Learning Objectives:

- Understand NLP applications in finance and mobile money
- Learn how chatbots process user intent
- Explore spam filtering with pattern recognition

Everyday NLP Applications (Cont.)

1. Mobile Money & Finance

Budgeting apps that automatically read your M-Pesa or bank SMS messages to calculate expenses.
NLP extracts the **amount**, **date**, and **recipient** from raw text messages.

2. Chatbots & Virtual Assistants

Intelligent bots that instantly answer customer questions on platforms like WhatsApp.
Models understand user **intent** (e.g., "Check balance") to provide relevant, automated support.

Everyday NLP Applications (Cont.)

3. Chatbots

Instant support from airlines or internet providers.

The model reads your question, understands **intent** (e.g., "Check balance"), and replies instantly.

4. Spam Detection

How Gmail filters "You have won $1,000,000" emails.

NLP models recognize **suspicious patterns** (ALL CAPS, too many $ signs, or urgent language).

Application 4: Machine Translation

Real-Time Language Solutions

Tools like Google Translate use advanced NLP to translate text from English to Swahili, French, or Kinyarwanda in real-time.

It doesn't just translate word-by-word; it translates the ***meaning*** of the whole sentence so the grammar makes sense.

The Main NLP Tasks

What specific jobs can we train our AI to do?

KNOWLEDGE CHECK

Module 2: Core Applications

1. How does NLP help budgeting apps read M-Pesa messages?

2. What does a chatbot need to understand to reply correctly?

3. What patterns does NLP look for when detecting spam emails?

4. How does machine translation differ from word-by-word translation?

Take a moment to reflect on these questions before moving on.

MODULE 3

The NLP Pipeline

Learning Objectives:

- Understand data ingestion and text cleaning
- Learn tokenization, stemming, and vectorization
- Build and evaluate NLP models step by step

Task 1: Text Classification

Assigning categories to text

This process is heavily used for routing customer support tickets to the right department automatically.

Input: *"The new Jumia delivery system is very fast."*
Output Category: *Logistics / Delivery*

Task 2: Sentiment Analysis

Determining Emotional Tone

Identifying the feeling behind a series of words as **Positive**, **Negative**, or **Neutral**.

Business Application:

Companies use this to analyze thousands of Twitter mentions to gauge public reception of new product launches.

NLP Tasks

Task 3: Named Entity Recognition

Automatically identifying specific nouns in text like Names, Organizations, Locations, and Dates.

*"Yesterday, **Amina** traveled to **Lagos** to visit **Flutterwave**."*

The AI knows **Amina** is a Person, **Lagos** is a City, and **Flutterwave** is a Company.

Task 4: Text Summarization

Taking a long document and shrinking it down into a few sentences.

Extractive Summarization

Copy-pasting the most important sentences directly from the source text.

Best for: Quick highlights and news snippets.

Abstractive Summarization

Writes a brand-new summary using its own words.

Task 5: Question Answering (Q&A)

Extracting Answers from Documents

You give the AI a document (like a **PDF of your country's constitution**), and you ask it a question in plain English.

How it works:

The AI reads the document, finds the exact paragraph, and extracts the answer for you. This is the technology behind **advanced search engines**.

The Text NLP Pipeline

Step-by-step Journey

How we turn messy human words into computer logic.

Step 1: Data Ingestion

Gathering the Foundation

Before we can do NLP, we need text! We call a large collection of text a **Corpus**.

Common methods to "ingest" data:

- Web-scraping news articles
- Downloading datasets from the internet
- Connecting to the Twitter API
- Creating your own dataset from scratch

The NLP Pipeline: Steps 2 & 3

Step 2: Validation & EDA

Exploratory Data Analysis

We look at our data before we touch it to avoid headaches later.

Key Questions:

- Are the texts too long?
- Do we have missing data?
- Multiple languages mixed?

Step 3: Text Cleaning

Pre-Processing

Machines see capitalization and punctuation as different words.

"Nairobi", "NAIROBI", "Nairobi!"
= 3 Different Words to Machines

- Remove punctuation & symbols
- Strip links & HTML tags
- **Convert to lowercase**

Step 4: Tokenization

Breaking Down Sentences

The computer cannot read a whole sentence at once. We must chop the sentence into smaller pieces called **Tokens**.

Example

Sentence: `"The food is great"`

Tokens: `["The", "food", "is", "great"]`

Step 5: Stop Words Removal

Filtering Out "Glue" Words

Human language contains "glue" words that don't add much actual meaning.

Examples:

`"the", "is", "at", "which", "on"`

We delete these **Stop Words** so the computer can focus purely on the **important words** like "food" and "great".

Step 6: Stemming & Lemmatization

Normalizing Word Variations

Words have many forms: **run, running, ran**. To a computer, we want these all to be treated as the exact same root word.

Stemming

Brutally chops off the end of words (e.g., "Running" -> "Run").

Lemmatization

Uses a dictionary to safely shrink a word to its base root (e.g., "Better" -> "Good").

Step 7 & 8: Finalizing the NLP Pipeline

Step 7: Vectorization

Turning text into math using algorithms like **TF-IDF** (Term Frequency-Inverse Document Frequency).
functionsCounts word occurrences and scores their importance.
grid_onCreates a giant matrix of numbers for the computer to read.

Step 8: Model Training

Feeding vectorized data into Machine Learning algorithms to learn patterns.
psychologyThe model identifies relationships between numbers and labels.
rocket_launchReady for evaluation and real-world predictions.

The model learns patterns (e.g., "When I see the numbers for 'terrible service', I predict a Negative review"). We then test it on new data to see its **Accuracy score**.

Python Libraries: NLTK

The Origin of Python NLP

The **Natural Language Toolkit (NLTK)** is a foundational library, ideal for learning the core concepts of language processing.
Excellent for beginners and academic research.
Powerful tools for tokenization and finding Stop Words.
Basic rule-based sentiment analysis via tools like VADER.

KNOWLEDGE CHECK

Module 3: The NLP Pipeline

1. List the 8 steps of the NLP pipeline in order.

2. What is tokenization and why is it necessary?

3. Explain the difference between stemming and lemmatization.

4. What does TF-IDF measure, and why is it useful for vectorization?

Take a moment to reflect on these questions before moving on.

MODULE 4

Tools & Libraries

Learning Objectives:

- Get hands-on with spaCy, NLTK, and Scikit-Learn
- Set up Google Colab for NLP experiments
- Explore audio processing with Librosa

Library 2: spaCy

While NLTK is great for learning, spaCy is designed for production use, offering exceptional speed and efficiency for processing massive text datasets.
High Performance: Optimized for speed and large-scale data.
Enterprise Ready: Trusted by big tech for real-world applications.
Advanced NER: Exceptional at Named Entity Recognition.

Library 3: Scikit-Learn

Machine Learning & Predictive Modeling

Scikit-Learn is the ultimate Machine Learning library for Python, providing simple and efficient tools for predictive data analysis.
Build prediction models such as Naive Bayes or Random Forest.
Calculate Accuracy scores to evaluate model performance
Standard tool for the data science workflow after text cleaning.

The Audio NLP Pipeline

Step-by-step Journey

How we turn audio into computer logic.

Why Audio NLP?

Empowering Spoken Communication

In many parts of Africa, literacy rates vary, and many local languages are primarily spoken rather than written.

Speech-to-Text (STT) allows users to interact with technology naturally.

If a farmer can speak into a USSD app instead of navigating complex text menus, the technology becomes accessible to everyone.

Audio Data Collection

Just like text needs a Corpus, Voice AI needs thousands of hours of audio recordings.

Sources

We collect audio from radio broadcasts, WhatsApp voice note donations, parliamentary speeches etc.

Diversity is Key

An AI trained only on Tanzania swahili accents will fail in Congo. We must collect diverse accents, ages, and genders to ensure the AI works for everyone.

The Challenge: Audio is Noisy

Real-world audio is incredibly messy compared to text.

"If you record a voice note at a busy public transport stage. The microphone picks up your voice, but also keke engines, overlapping conversations, wind, and hawkers in the background."

Raw, noisy audio confuses AI models. Cleaning the audio is a mandatory first step before processing.

Audio Cleaning (Noise Reduction)

Audio cleaning is the equivalent of removing "Stop Words" in text.

Silence Removal: Automatically cutting out the dead air where nobody is speaking to save computer processing time.

Background Noise Filtering: Using algorithms to separate human voice frequencies from static hums, wind, or street noise.

Normalization: Making sure the audio isn't too quiet or too loud.

Audio Preprocessing: Sampling

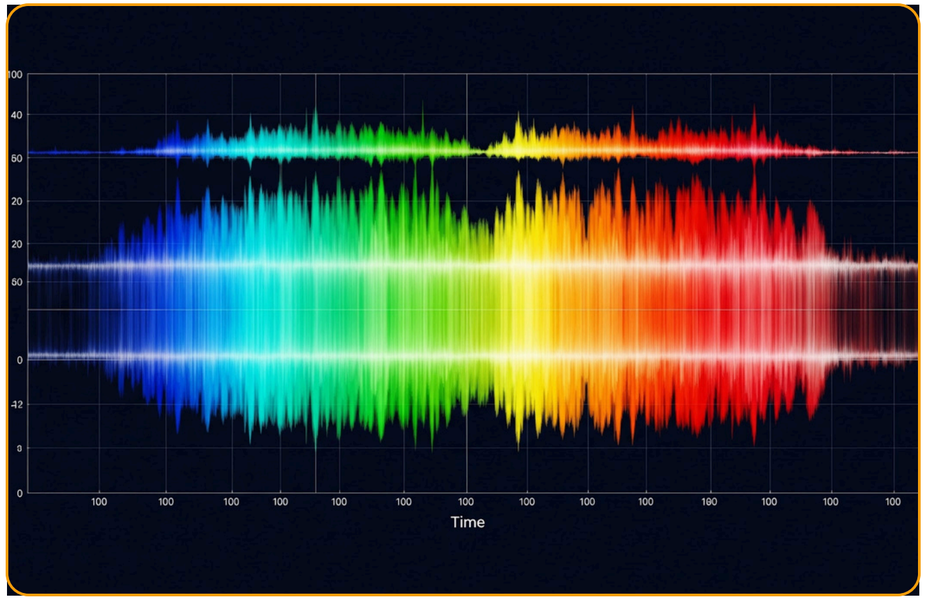
Computers don't have ears. They process sound as numbers.
To do this, they take snapshots of the sound wave thousands of times per second.
Sample Rate: This is what we call the frequency of these snapshots.
16,000 Hz: A standard sample rate for Voice AI, meaning the computer measures the audio wave 16,000 times every single second!

Audio Preprocessing: Spectrograms

How does an AI actually "read" sound? We turn the audio into a picture!

Using a mathematical trick called a **Fourier Transform**, we create a **Spectrogram** . It is a visual heatmap of sound frequencies over time.

The model then looks at this image to figure out what words were spoken.



Example of a Spectrogram Image

Audio NLP Libraries: Librosa

When working with Voice AI, Python developers use a library called Librosa.

It provides a comprehensive toolset for audio analysis and preprocessing:

- Load and manipulate **MP3 or WAV** files
- Extract **sample rates** accurately
- Filter out unwanted **background noise**
- Generate high-quality **Spectrogram images** for AI training

KNOWLEDGE CHECK

Module 4: Tools & Libraries

1. What is the difference between NLTK and spaCy in terms of use case?

2. Why is Google Colab useful for training AI models?

3. What is a spectrogram and how does it help Voice AI?

4. Name two reasons why audio data must be cleaned before processing.

Take a moment to reflect on these questions before moving on.

MODULE 5

Transfer Learning & BERT

Learning Objectives:

- Understand transfer learning and why it matters
- Explore BERT and HuggingFace Transformers
- Fine-tune pre-trained models for African languages

Google Colab & AI Training

Google Colab

Training AI requires very powerful, expensive laptops with GPUs (Graphics Cards).
Google Colab gives you a free, super-fast cloud computer right inside your web browser.
You write Python code in your browser, and Google's servers do the heavy lifting
Access the google colab here:

Data Annotation & Ethics

Data Annotation

AI is not magic; it only knows what humans teach it.

To teach an AI to understand text, human workers must sit down and manually label thousands of sentences.

This manual labeling is called **Data Annotation**.

"Garbage In, Garbage Out"

The Human Factor

If the humans labeling the data do a poor job, or disagree with each other, the dataset will be "Garbage."

The Impact on AI

If you train a multi-million dollar AI on a garbage dataset, you will get a garbage model.

Data quality is the most important part of NLP.

Bias and Sustainability in AI

Bias & Underrepresentation

AI models inherit training data biases, often skewed towards Western sources.

Example: An AI might know "Cheeseburger" but fail to recognize "Pap" or "Jollof".

Action: We must consciously build inclusive datasets!

Green AI: Environmental Cost

Training LLMs consumes vast resources:

- Vast electricity consumption
- Significant water usage
- High carbon footprint

Practice "Green AI" make smaller, efficient models rather than blindly burning energy.

Transfer Learning & BERT

The modern revolution in Natural Language Processing.

Deep Learning is shifting from training models from scratch to fine-tuning massive pre-trained architectures.

BERT: Bidirectional Encoder Representations from Transformers

What is Transfer Learning?

The Traditional Approach

Traditionally, every time we wanted to do a new task, we had to build a model from absolute zero. This took too much time and data.

The Transfer Learning Revolution

Transfer Learning is taking a model that has already been trained on one massive task (like reading all of Wikipedia), and "transferring" that knowledge to a new, smaller task.

Why is Transfer Learning amazing?

Saves Time

You don't need a super-computer running for 3 months. You can adapt a pre-trained model on your laptop in an hour.

Needs Less Data

Because the model already "knows" English grammar, you only need a few hundred examples for your specific task, instead of millions!

Fine-Tuning & BERT

What is Fine-Tuning?

The process of adapting a pre-trained model to your specific use-case.
We freeze the "brain" and only train the final layer for specific predictions like "Positive" or "Negative".

Enter BERT

Bidirectional Encoder Representations from Transformers is a famous pre-trained model created by Google.
Before BERT, models read text strictly from left-to-right. BERT is "Bidirectional" it reads the whole sentence at once, looking both left and right to understand context.

KNOWLEDGE CHECK

Module 5: Transfer Learning & BERT

1. What is transfer learning and why is it a 'revolution'?

2. How does BERT differ from older models that read left-to-right?

3. What are contextual embeddings? Give an example.

4. Name two Afro-centric language models and their purpose.

Take a moment to reflect on these questions before moving on.

MODULE 6

Responsible AI

Learning Objectives:

- Recognize bias in NLP datasets and models
- Build inclusive, representative training data
- Apply Green AI principles for sustainability

Contextual Embeddings

Older models assigned a fixed mathematical value to each word. **BERT** generates **contextual embeddings**, meaning the value changes based on surrounding words.

Example: "Crane" (Nature)

"The **crane** flew over the lake."

In this context, the model identifies the word as a bird and assigns a specific mathematical representation.

Example: "Crane" (Industry)

"The **crane** lifted the steel beam."

The model recognizes the construction site context, giving "crane" a completely different numerical value.

Your Next Step: Hugging Face

Hugging Face is the "GitHub of Machine Learning"

It is the central hub where you can download state-of-the-art pre-trained models like **BERT** for free, explore datasets, and share your own AI projects.

Action Item:

Go to **HuggingFace.co**, create your free account, and prepare to code!

Bridging African Tongues

Afro-centric Models

Transfer learning is critical for low-resource African languages where massive text datasets don't exist.

AfriBERTa

Trained specifically on 11 African languages.

AfroLM

Covers 23 languages using self-active learning.

Masakhane

Leading the participatory research movement for African Translation.

Action Guide: Building a Local Corpus

1 Identify Sources

Find where the language lives: Social Media (X/Reddit), Messaging (WhatsApp), Music Lyrics, or Radio Transcripts.

2 Data Extraction

Pull text using Python (BeautifulSoup, Scrapy) or export chat logs as .txt files for processing.

3 Cleaning & Privacy

Remove PII, standardize tricky dialect spellings, and use Regex to strip HTML or broken emojis.

4 Annotation & Labeling

Use tools like Label Studio or Doccano. Create translation pairs (English vs. Dialect) or assign sentiment labels with native speakers.

5 Open Source & Share

Upload your CSV to Hugging Face. Write a "Data Card" documenting sources and biases to help the community.

Pro-Tip

Even a high-quality dataset of 1,000 to 5,000 sentences is enough to fine-tune pre-trained models using Transfer Learning!

Extra Reading List Recommendation

- Advanced NLP, Graham Neubig http://www.phontron.com/class/anlp2022/
- Advanced NLP, Mohit Iyyer https://people.cs.umass.edu/~miyyer/cs685/
- NLP with Deep Learning, Chris Manning, http://web.stanford.edu/class/cs224n/
- Understanding Large Language Models, Danqi Chen https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
- Natural Language Processing, Greg Durrett https://www.cs.utexas.edu/~gdurrett/courses/online-course/materials.html
- Large Language Models: https://stanford-cs324.github.io/winter2022/
- Natural Language Processing at UMBC, https://laramartin.net/NLP-class/
- Computational Ethics in NLP, https://demo.clab.cs.cmu.edu/ethical_nlp/
- Self-supervised models, CS 601.471/671: Self-supervised Models (jhu.edu)
- WING.NUS Large Language Models, https://wing-nus.github.io/cs6101

KNOWLEDGE CHECK

Module 6: Responsible AI

1. What does 'Garbage In, Garbage Out' mean for AI training?

2. Give an example of bias in NLP models.

3. What is Green AI and why does it matter?

4. Describe 3 steps for building a local language corpus.

Take a moment to reflect on these questions before moving on.

COURSE COMPLETE

Congratulations!

You have completed the NLP & African Language Models Pre-course.
You now have the foundation to build NLP pipelines, apply transfer learning, and create inclusive AI for African languages.

What you learned:

NLP Fundamentals | Core Applications | The NLP Pipeline
Tools & Libraries | Transfer Learning & BERT | Responsible AI

Instructors: Abigail Oppong & Loise Mwarangu | DSA 2026