MASTER OF SCIENCE OF INFORMATION SYSTEMS
INTELLIGENT SYSTEMS

Natural Language Processing

Sentiment Analysis -

AfriSenti Twitter Sentiment dataset

STUDENTS (GROUP2)	AINEDEMBE DENIS	2024-M132-23999
	MUSINGUZI BENSON	2024-M132-23947
LECTURER	Dr. Sibitenda Harriet	

About the AfriSenti Twitter Sentiment Dataset

- The dataset contains 110,000+ annotated tweets across 14 African languages; Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, Yoruba

- Tweets are labeled by three annotators into three sentiment categories: Positive, Negative, Neutral

- All tweets are anonymized (@user) and URLs removed to protect user privacy

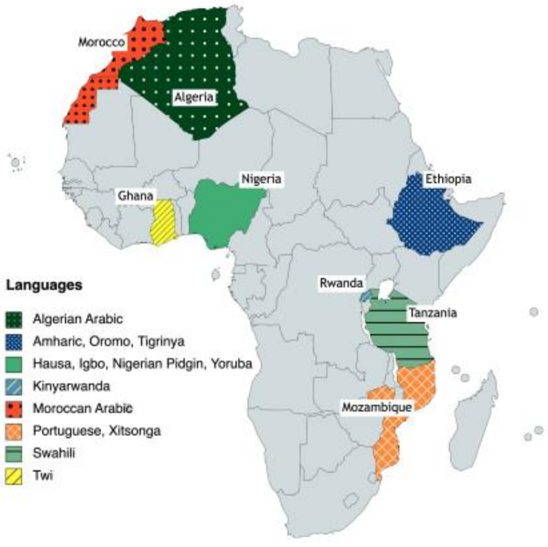
- Designed for Monolingual and multilingual and sentiment analysis, supporting research on African languages which are underrepresented in NLP.

Why this dataset matters:

- African languages make up 30% of the world's languages, yet lack NLP datasets (UNESCO) - AfriSenti fills this gap

AFRISENTI
SemEval 2023

AfriSenti
Sentiment Analysis dataset for 14 African languages



References:

- Codalab: https://codalab.lisn.upsaclay.fr/competitions/7320 ,
- Hugging Face: https://huggingface.co/datasets/HausaNLP/AfriSenti-Twitter
- ACL Anthology: https://aclanthology.org/2023.emnlp-main.862.pdf

Project Objectives

Main Objective

- To build and evaluate a multilingual sentiment analysis model that classifies African-language tweets into positive, negative, or neutral sentiment.

Specific Objectives

- **Explore the dataset:** Analyze language distribution, tweet length, and sentiment label imbalance.
- **Preprocess the text:** Clean tweets, remove URLs, normalize slang, and tokenize using mBERT or XLM-RoBERTa.
- **Model Development:** Fine-tune XLM-RoBERTa for 3-class sentiment classification and compare with an LSTM baseline.
- **Model Training:** Use early stopping and gradient clipping for 3–5 epochs to prevent overfitting.
- **Evaluation:** Measure Accuracy, Macro-F1, ROC-AUC, and produce a confusion matrix. Include example predictions and attention visualization.
- **Ablation Studies:** Vary batch size, learning rate, and sequence length to observe performance changes.
- **Cross-Lingual Testing:** Train on one language (e.g., Swahili) - test on another (e.g., Amharic) to assess the model’s multilingual transfer ability.

Real-Life Applicability

1. Social Media Monitoring

Governments, NGOs, and research institutions can track public mood on elections, health, crises, and social issues using multilingual sentiment trends.

2. Customer & Brand Analysis

Businesses can monitor consumer reactions across languages to improve products and services.

3. Hate Speech & Online Safety

Sentiment analysis is a key component in detecting negative or harmful content, especially in multilingual online spaces.

4. Policy & Public Opinion Research

Analysts can understand public attitudes in African countries/languages that are often ignored due to lack of datasets.

5. Enhancing African NLP Tools

AfriSenti supports the development of digital tools for African languages, promoting inclusive and equitable AI technology

6. Cross-Lingual AI Applications

Your cross-lingual experiments contribute to systems that work even when certain languages lack large datasets.

Loading AfriSenti-Twitter Dataset: Listing Available Language Configurations

```
LISTING AVAILABLE LANGUAGE CONFIGURATIONS FOR AfriSenti-Twitter DATASET
______________________________________________________________________________________________________________________________
Available language configs: ['amh', 'hau', 'ibo', 'arq', 'ary', 'yor', 'por', 'twi', 'tso', 'tir', 'orm', 'pcm', 'kin', 'swa']

Loading Amharic (amh) with all splits
----------------------------------------
DatasetDict({
    train: Dataset({
        features: ['tweet', 'label'],
        num_rows: 5984
    })
    validation: Dataset({
        features: ['tweet', 'label'],
        num_rows: 1497
    })
    test: Dataset({
        features: ['tweet', 'label'],
        num_rows: 1999
    })
})

Loading a single split /train only for Amharic (amh):
------------------------------------------------------------------------------------------
{'tweet': 'Tesfaye ለካስ ጭብል ለብሰሽ የፕሮፌሰርን ፎቶ ለጥፈክ እልም ያልክ ባዳ ነክ እፈር ትንሽ', 'label': 2}
```

```
Language Code & Language Name Mapping:
______________________________________________
  amh  - Amharic
  arq  - Algerian Arabic
  ary  - Moroccan Arabic
  hau  - Hausa
  ibo  - Igbo
  kin  - Kinyarwanda
  orm  - Oromo
  pcm  - Nigerian Pidgin
  por  - Portuguese
  swa  - Swahili
  tir  - Tigrinya
  tso  - Tsonga
  twi  - Twi
  yor  - Yoruba
```

Loading the AfriSenti Twitter Sentiment dataset

```
DATASET SAMPLE: One Example from Each Language

____________________________________________________________________________

Loading one sample from each of 14 languages...



----------------------------------------------------------------------------
Lang.                                                    Tweet Sentiment
  amh Tesfaye ለካስ ጭብል ለብሰሽ የፕሮፌሰርን ፎቶ ለጥፈክ እልም ያልክ ባዳ ነክ...  Positive
  arq @user على حسب موقعك يبدو أنك صاحب نظرة ثاقبة .يخي ...  Positive
  ary hhhhhhhhhhhhhhhhhhhhhh ana ga3ma sma3tt ach kant k...   Neutral
  hau @user Da kudin da Arewa babu wani abin azo agani d...  Positive
  ibo        Nna Ike Gwuru ooo. 😂 https://t.co/NDS7juFBGd  Positive
  kin @user @user @user @user @user @user @user Hhhhhh n...  Positive
  orm @user Waa’ee mana waaqeffanaa ilaalcha keessa galc...   Neutral
  pcm yeah the guy wants to trend dat was why e join n...  Positive
  por Pedi uma resposta a Deus, ele deu me. Estou muito ...  Positive
  swa Kwani tanesco wanakataga umeme makusudinadhani kun...  Positive
  tir @user @user @user @user እንታይ ካብ ጎንደር ዝወረስዎን ድኣ : ባ...  Positive
  tso             @user Loku u navela Ku tissunga, tissungue  Positive
  twi                              kako be shark but wo ti ewu  Positive
  yor Ìwọ ikú òpònú abaradúdú wọ, o ò ṣe é 're o. O d'ór...  Positive
----------------------------------------------------------------------------

Total languages shown: 14/14
____________________________________________________________________________
```

Collect one example from each language

Label mapping: 0=negative, 1=neutral, 2=positive

Truncate tweet to 50 characters for better display

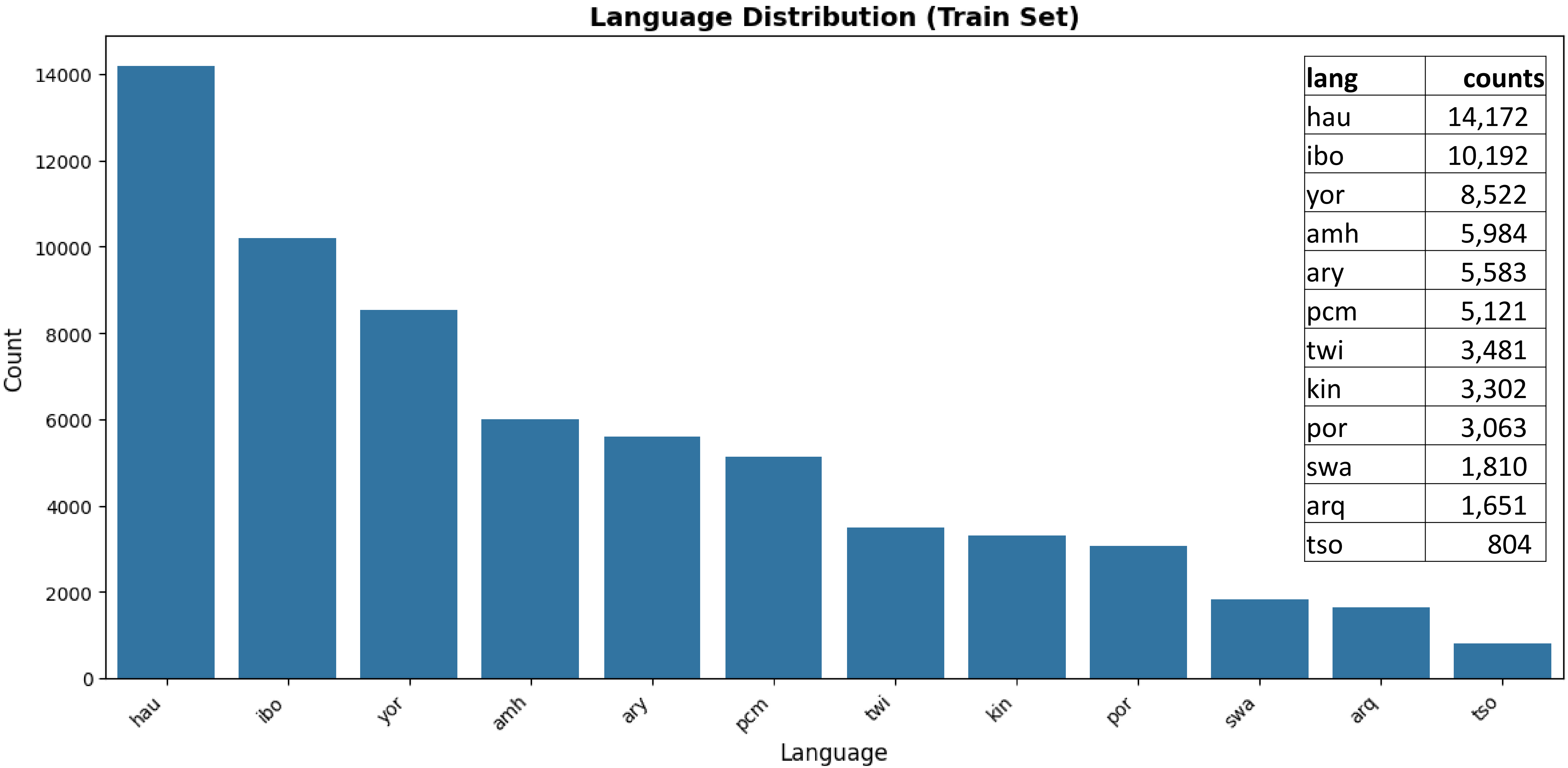
AfriSenti Twitter dataset: Total tweets across all languages

```
Language configs: ['amh', 'hau', 'ibo', 'arq', 'ary', 'yor', 'por', 'twi', 'tso', 'tir', 'orm', 'pcm', 'kin', 'swa']
      train  validation  test  total
lang
amh    5984        1497  1999   9480
hau   14172        2677  5303  22152
ibo   10192        1841  3682  15715
arq    1651         414   958   3023
ary    5583         494  2961   9038
yor    8522        2090  4515  15127
por    3063         767  3662   7492
twi    3481         388   949   4818
tso     804         203   254   1261
tir       0         398  2000   2398
orm       0         396  2096   2492
pcm    5121        1281  4154  10556
kin    3302         827  1026   5155
swa    1810         453   748   3011

Total tweets across all languages: 111718
```

Initial Data Exploration: Load train data from all languages

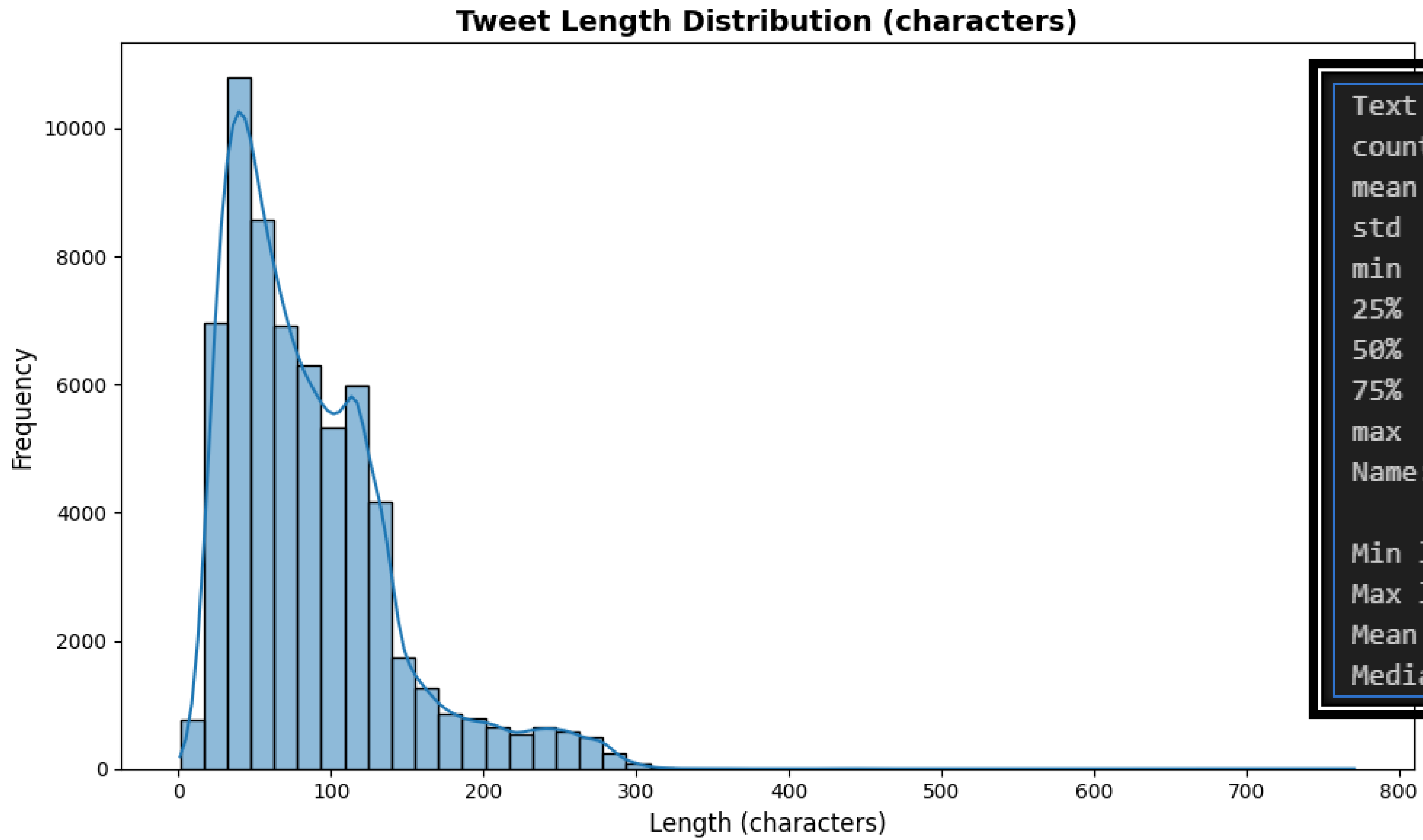
```
Dataset Loading Summary
----------------------------------------------------------------------
Language Code   Language Name  Train Samples         Status
          amh         Amharic           5984         Loaded
          hau           Hausa          14172         Loaded
          ibo            Igbo          10192         Loaded
          arq Algerian Arabic           1651         Loaded
          ary Moroccan Arabic           5583         Loaded
          yor          Yoruba           8522         Loaded
          por      Portuguese           3063         Loaded
          twi             Twi           3481         Loaded
          tso          Tsonga            804         Loaded
          tir        Tigrinya              0 No train split
          orm           Oromo              0 No train split
          pcm Nigerian Pidgin           5121         Loaded
          kin     Kinyarwanda           3302         Loaded
          swa         Swahili           1810         Loaded
----------------------------------------------------------------------

Total train samples loaded: 63,685

First few rows:
  lang lang_name                                             tweet  label label_text
0  amh   Amharic  Tesfaye ለካስ ጭብል ለብሰሽ የፕሮፌሰርን ፎቶ ለጥፈክ እልም ያልክ ባ...      2   positive
1  amh   Amharic  ይሄው ነው አይደል የእውቀትሽ ጥግ....በሰሚ ሰሚ ከምትናገረ ለምን ታሪክ...      2   positive
2  amh   Amharic                  ዘገበ ይባላል? ሌላ የሚባል ነገር ካለ አንተው ንገረን!      2   positive
3  amh   Amharic  ?? ድሮ በዘመነ ኮዳክ ፎቶ ቤት ፍላሹ ፏ ሲል አይናችን ተጨፍኖ እንዳይወ...      2   positive
4  amh   Amharic                                    ዠልጥ?? ???? ገገማ      2   positive
```

Initial Data Exploration: Language Distribution on the Train Set



lang	counts
hau	14,172
ibo	10,192
yor	8,522
amh	5,984
ary	5,583
pcm	5,121
twi	3,481
kin	3,302
por	3,063
swa	1,810
arq	1,651
tso	804

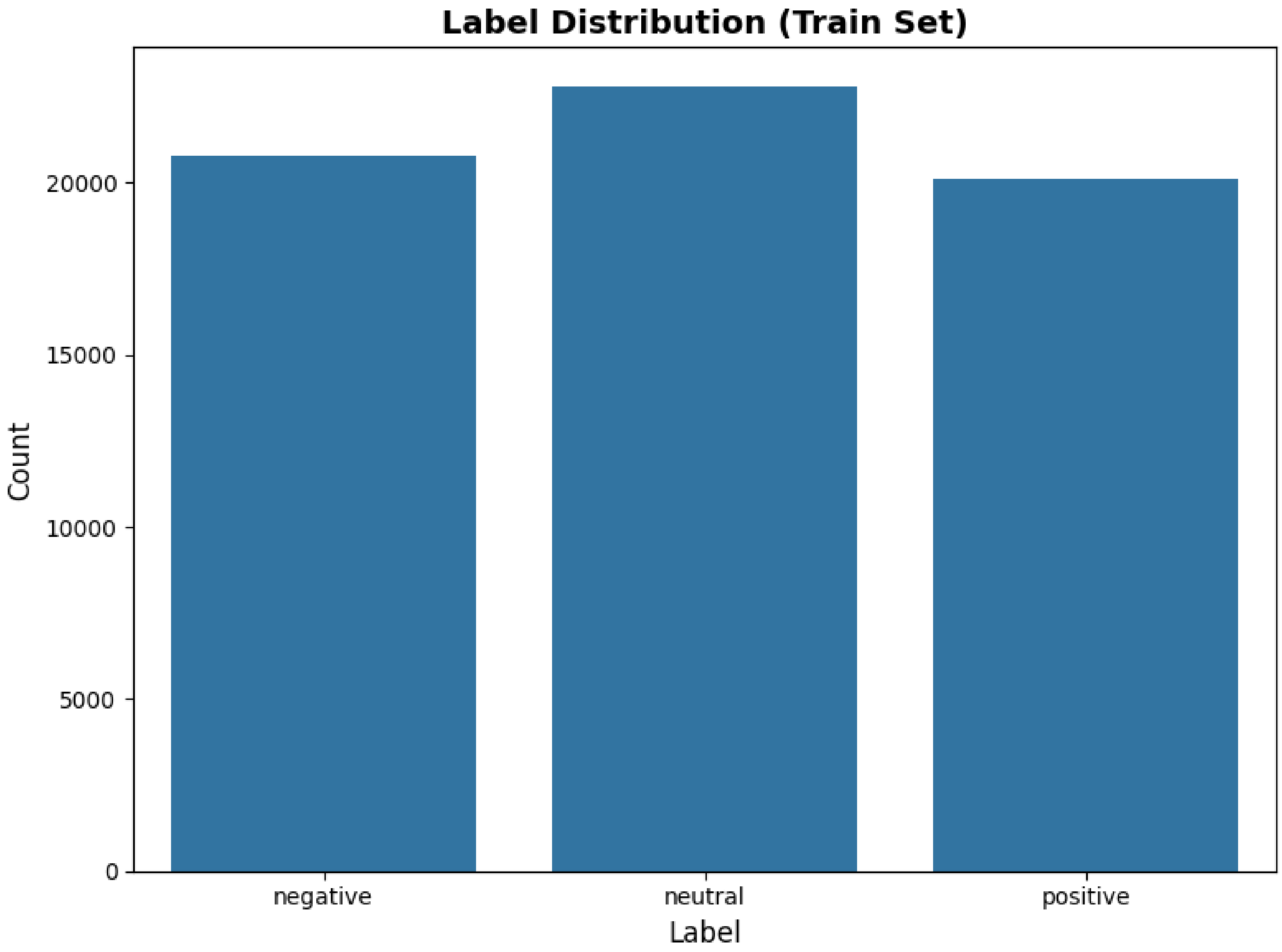
Initial Data Exploration: Text length Distribution the Train Set




```
Text length statistics:
count    63685.000000
mean        85.523781
std         55.263220
min          1.000000
25%         44.000000
50%         73.000000
75%        115.000000
max        771.000000
Name: text_len, dtype: float64

Min length: 1
Max length: 771
Mean length: 85.52
Median length: 73.00
```

The training tweets are mostly short (median 73 characters), with a right-skewed distribution and a few very long outliers.
The data is concise and typical of Twitter usage, making it suitable for transformer models with moderate sequence lengths

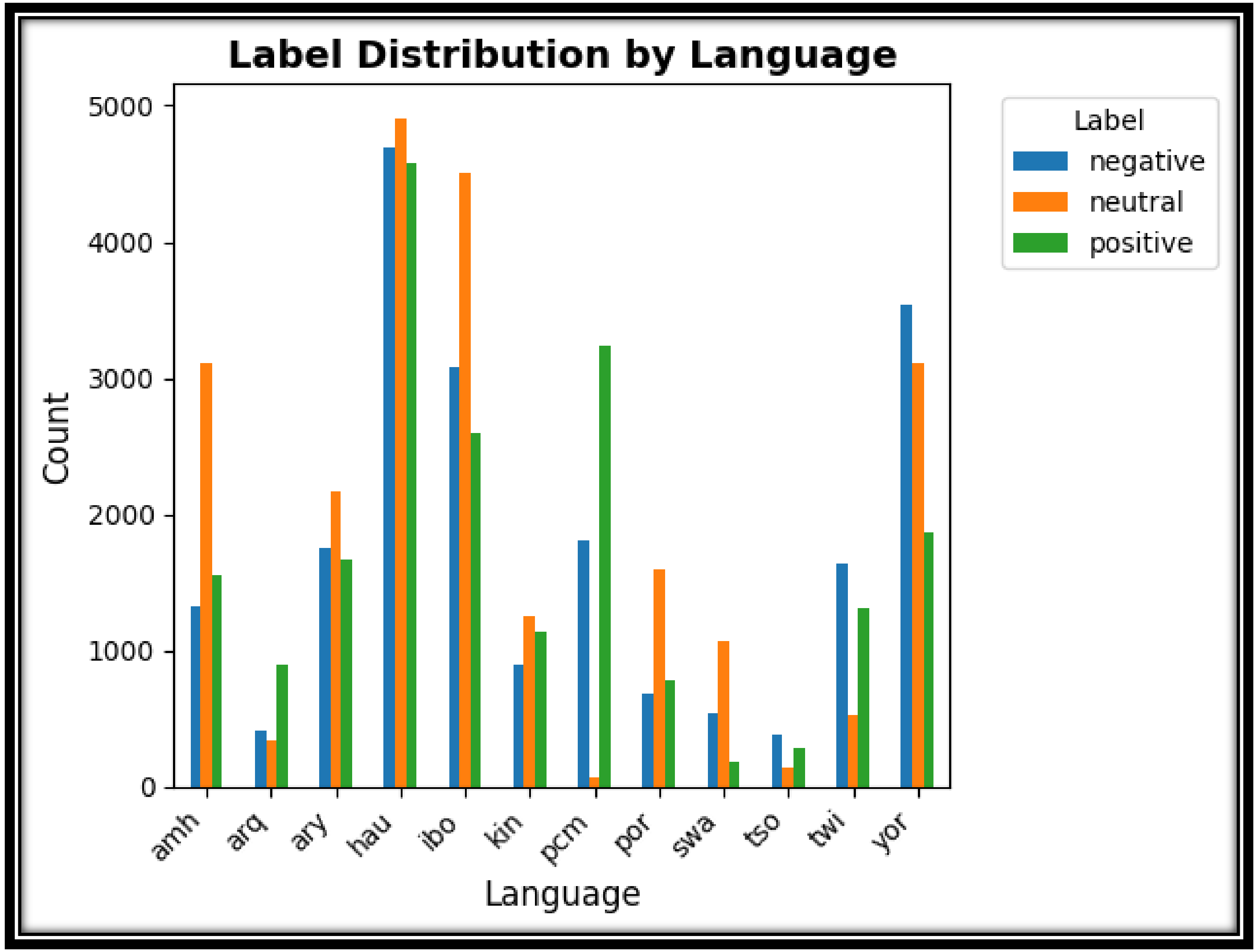
Initial Data Exploration: Overall Label Distribution




```
Label counts:
label_text
neutral      22794
negative     20783
positive     20108
Name: count, dtype: int64

Label percentages:
label_text
neutral      35.791788
negative     32.634058
positive     31.574154
Name: proportion, dtype: float64
```

The train set shows a well-balanced label distribution with 36% neutral, 33% negative, and 32% positive tweets. Thes observed balance supports stable model training without the need for class reweighting.

Initial Data Exploration: Label Distribution by Language



Label distribution by language:

label_text lang	negative	neutral	positive
amh	1332	3104	1548
arq	417	342	892
ary	1758	2161	1664
hau	4687	4912	4573
ibo	3084	4508	2600
kin	899	1257	1146
pcm	1808	72	3241
por	681	1600	782
swa	547	1072	191
tso	384	136	284
twi	1644	522	1315
yor	3542	3108	1872

- 1) Sentiment distribution varies significantly across languages

- 2) Shome languages like (Hausa, Igbo, and Yoruba) are balanced, others (e.g., Nigerian Pidgin and Twi) show extreme skewness in one sentiment class.

- 3) Potential challenges for model training and emphasizes the importance of multilingual transfer and careful evaluation

Preprocess text using multilingual tokenizers (mBERT, XLM-RoBERTa)

Load train, validation, and test datasets from all languages

```
Train samples: 63685
Validation samples: 13726
Test samples: 34307
```

Text preprocessing. Handled emojis, URLs, and slang normalization.

```
Text preprocessing completed

Sample preprocessed tweets:
  1. Tesfaye ለካስ ጭብል ለብሰሽ የፕሮፌሰርን ፎቶ ለጥፈክ እልም ያልክ ባዳ ነክ እፈር ትንሽ
  2. ይሄው ነው አይደል የእውቀትሽ ጥግ....በሰሚ ሰሚ ከምትናገሪ ለምን ታሪክ አታነቢም....ደሞ ራስሽን አታስገምቺ
  3. ዘገበ ይባላል? ሌላ የሚባል ነገር ካለ አንተዉ ንገረን!
```

- URLs
- Mentions <USER>
- Normalize whitespace
- Remove # but keep word

Preprocess text: Label mapping & tokenizing Tweets Using XLM-RoBERTa

```
Label mappings:
 - label2id: {'negative': 0, 'neutral': 1, 'positive': 2}
 - id2label: {0: 'negative', 1: 'neutral', 2: 'positive'}

Map: 100%|██████████| 63685/63685 [00:00<00:00, 288476.51 examples/s]
Map: 100%|██████████| 13726/13726 [00:00<00:00, 108999.68 examples/s]
Map: 100%|██████████| 34307/34307 [00:00<00:00, 240070.99 examples/s]

Label encoding completed
```

Label mapping:

string -> int
(0=negative, 1=neutral, 2=positive)

Tokenizing Tweets

Using multilingual tokenizer XLM-RoBERTa

```
Loading multilingual tokenizer: xlm-roberta-base
Tokenizer loaded successfully: xlm-roberta-base
  Vocabulary size: 250,002

Max sequence length: 128

Tokenizing datasets with multilingual tokenizer

Map: 100%|██████████| 63685/63685 [00:07<00:00, 9038.61 examples/s]
Map: 100%|██████████| 13726/13726 [00:01<00:00, 8075.26 examples/s]
Map: 100%|██████████| 34307/34307 [00:03<00:00, 8613.86 examples/s]

Tokenization completed using multilingual tokenizer
```

Modeling

Fine-tune XLM-RoBERTa for 3-class sentiment classification

(positive, neutral, negative).

Compare with

LSTM baseline Model

XLM-RoBERTa Transformer Model

```
Device set to: cpu
Loading XLM-RoBERTa model: xlm-roberta-base

  - Number of labels: 3
  - Label mappings: {'negative': 0, 'neutral': 1, 'positive': 2}

Some weights of XLMRobertaForSequenceClassification were not init
You should probably TRAIN this model on a down-stream task to be

Successfully loaded the xlm-roberta-base model
Total parameters: 278,045,955
```

Total parameters:
278,045,955

LSTM Baseline Model

```
LSTM Model Configuration:
  - Vocabulary size: 250,002
  - Embedding dimension: 128
  - Hidden dimension: 128
  - Number of labels: 3

LSTM model created
  - Embedding dim: 128
  - Hidden dim: 128
  - Total parameters: 32,265,219
```

Total parameters:
32,265,219

Comparison: XLM-RoBERTa vs LSTM Baseline

```
PARAMETER COMPARISON:
____________________________________________________________________________________
  - XLM-RoBERTa has 8.6x more parameters than LSTM Baseline
  - XLM-RoBERTa: 278,045,955 parameters (pre-trained, fine-tuned on AfriSenti)
  - LSTM Baseline: 32,265,219 parameters (trained from scratch on AfriSenti)
  - Difference: 245,780,736 parameters (761.8% more)

  Note: XLM-RoBERTa's larger parameter count reflects its pre-trained multilingual
        knowledge, while LSTM is a lighter baseline model trained only on this dataset.
```

Training Models - 5 epochs with Early stopping + Gradient clipping)

Training LSTM Baseline Model with early stopping

```
Training LSTM Baseline Model

____________________________________________________________
[LSTM] Epoch 1/3 | Train loss: 0.9572 | Val acc: 0.5762 | Val F1: 0.5683
  - New best F1: 0.5683, model saved!
[LSTM] Epoch 2/3 | Train loss: 0.7799 | Val acc: 0.6189 | Val F1: 0.6160
  - New best F1: 0.6160, model saved!
[LSTM] Epoch 3/3 | Train loss: 0.6447 | Val acc: 0.6282 | Val F1: 0.6262
  - New best F1: 0.6262, model saved!

LSTM training completed. Best F1: 0.6262
```

Training Models - 5 epochs with Early stopping + Gradient clipping)

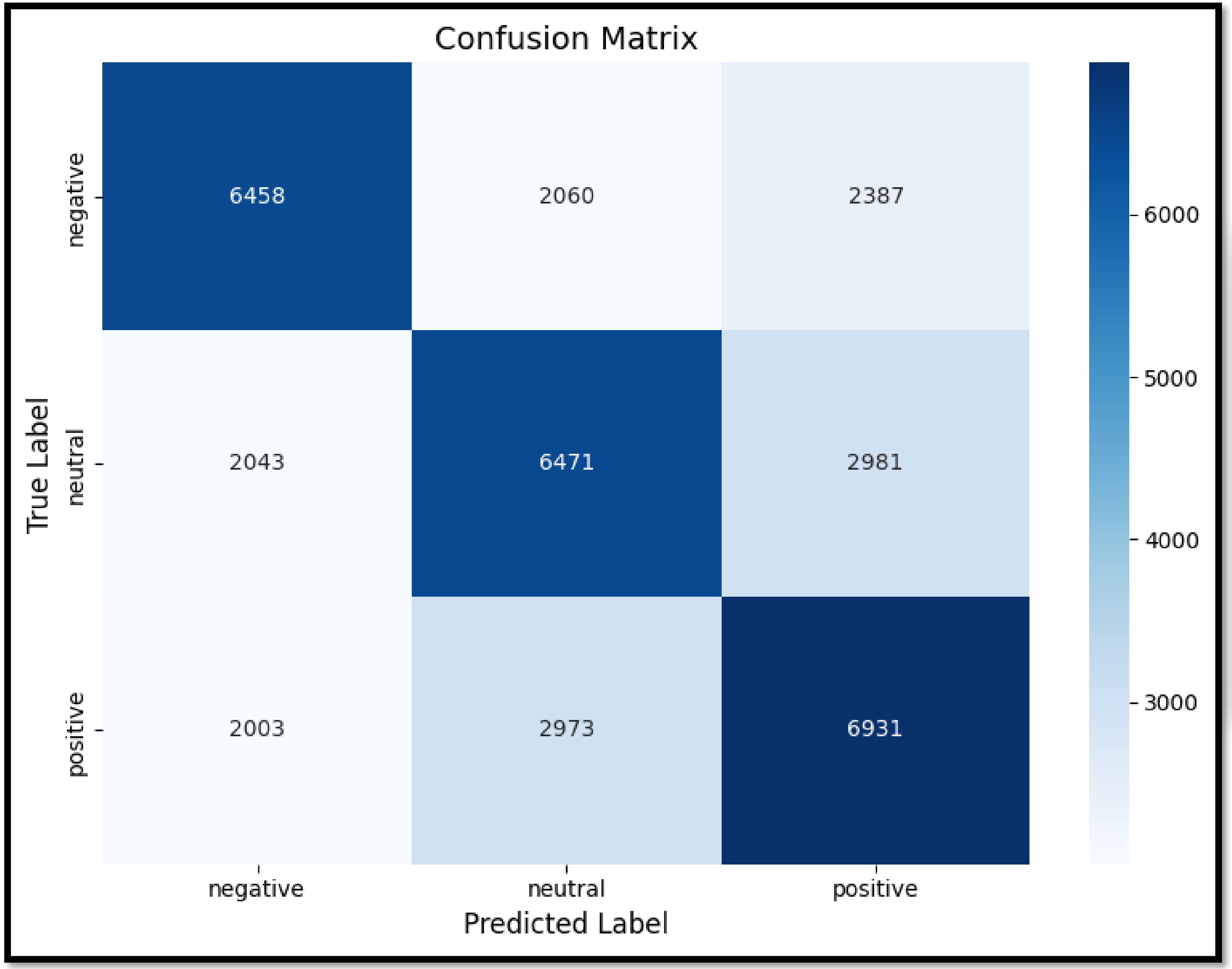
Training **XLM-RoBERTa** or AfriBERTa Model with early stopping

```
Training Transformer Model (xlm-roberta-base)

____________________________________________________________
------------------------------------------------------------
[TRANS] Epoch 1/3
  Train loss: 0.8997 | Val acc: 0.6241 | Val F1: 0.6242
  New best F1: 0.6242, model saved!
[TRANS] Epoch 2/3
  Train loss: 0.7241 | Val acc: 0.6689 | Val F1: 0.6688
  New best F1: 0.6688, model saved!
[TRANS] Epoch 3/3
  Train loss: 0.6122 | Val acc: 0.6792 | Val F1: 0.6795
  New best F1: 0.6795, model saved!
Transformer training completed. Best F1: 0.6795
Total training time: 91.3 minutes (1.52 hours)
```

Evaluation (F1, Accuracy, ROC-AUC, Confusion Matrix) + Predictions & Attention

Evaluating both LSTM and XLM-RoBERTa models on test data with comprehensive metrics, example predictions, and attention visualization.

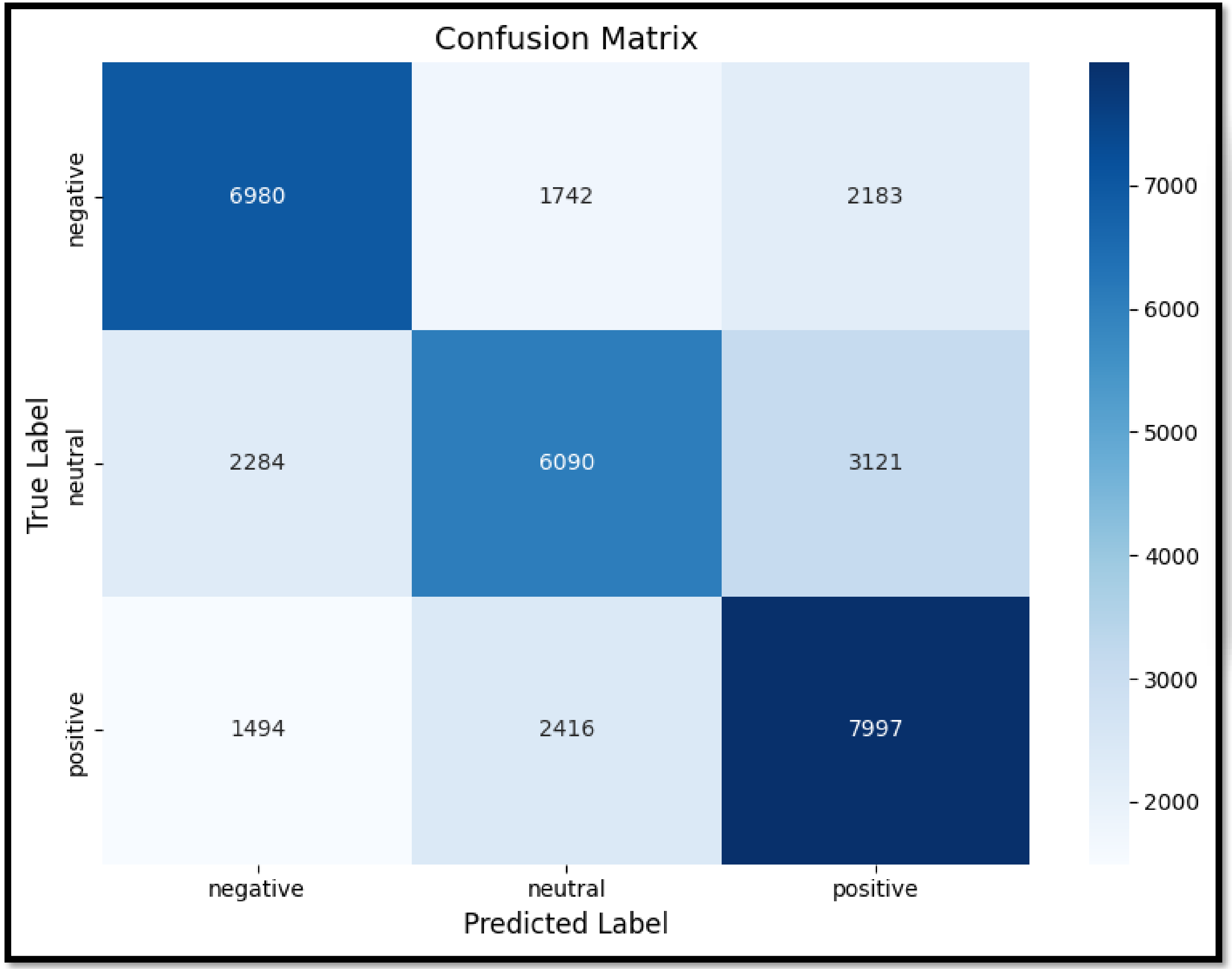
Evaluating LSTM Baseline Model on Test Set

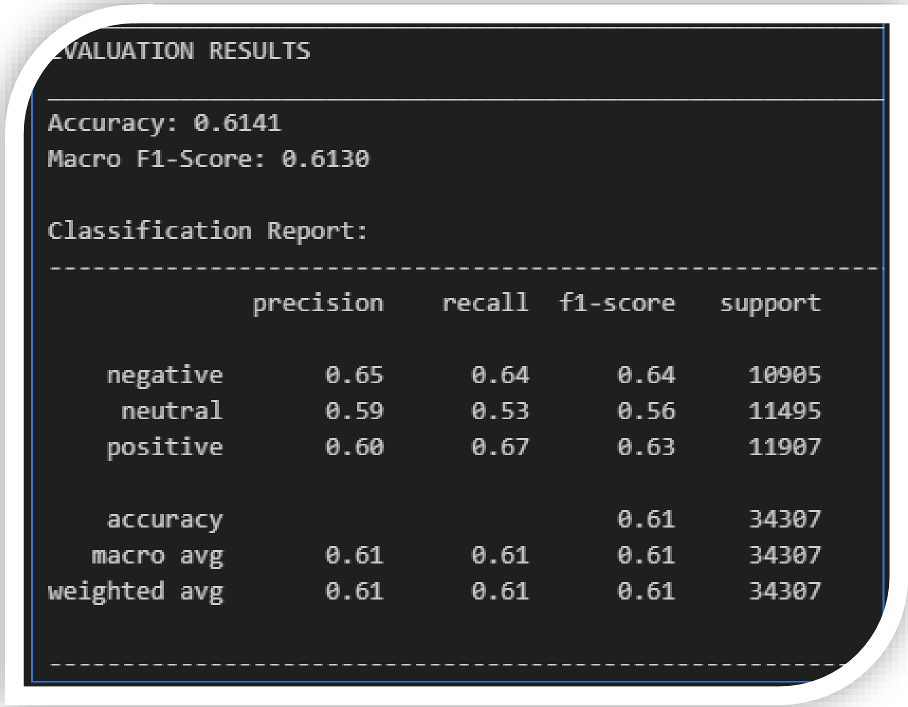



```
VALUATION RESULTS
________________________________________________________________
Accuracy: 0.5789
Macro F1-Score: 0.5796

Classification Report:
----------------------------------------------------------------
              precision    recall  f1-score   support

    negative       0.61      0.59      0.60     10905
     neutral       0.56      0.56      0.56     11495
    positive       0.56      0.58      0.57     11907

    accuracy                           0.58     34307
   macro avg       0.58      0.58      0.58     34307
weighted avg       0.58      0.58      0.58     34307

----------------------------------------------------------------
```

- Macro F1-Score: 0.5796
- Macro ROC-AUC: 0.7590

Evaluating XLM-ROBERTA Transformer Model on Test Set





EVALUATION RESULTS

Accuracy: 0.6141
Macro F1-Score: 0.6130

Classification Report:

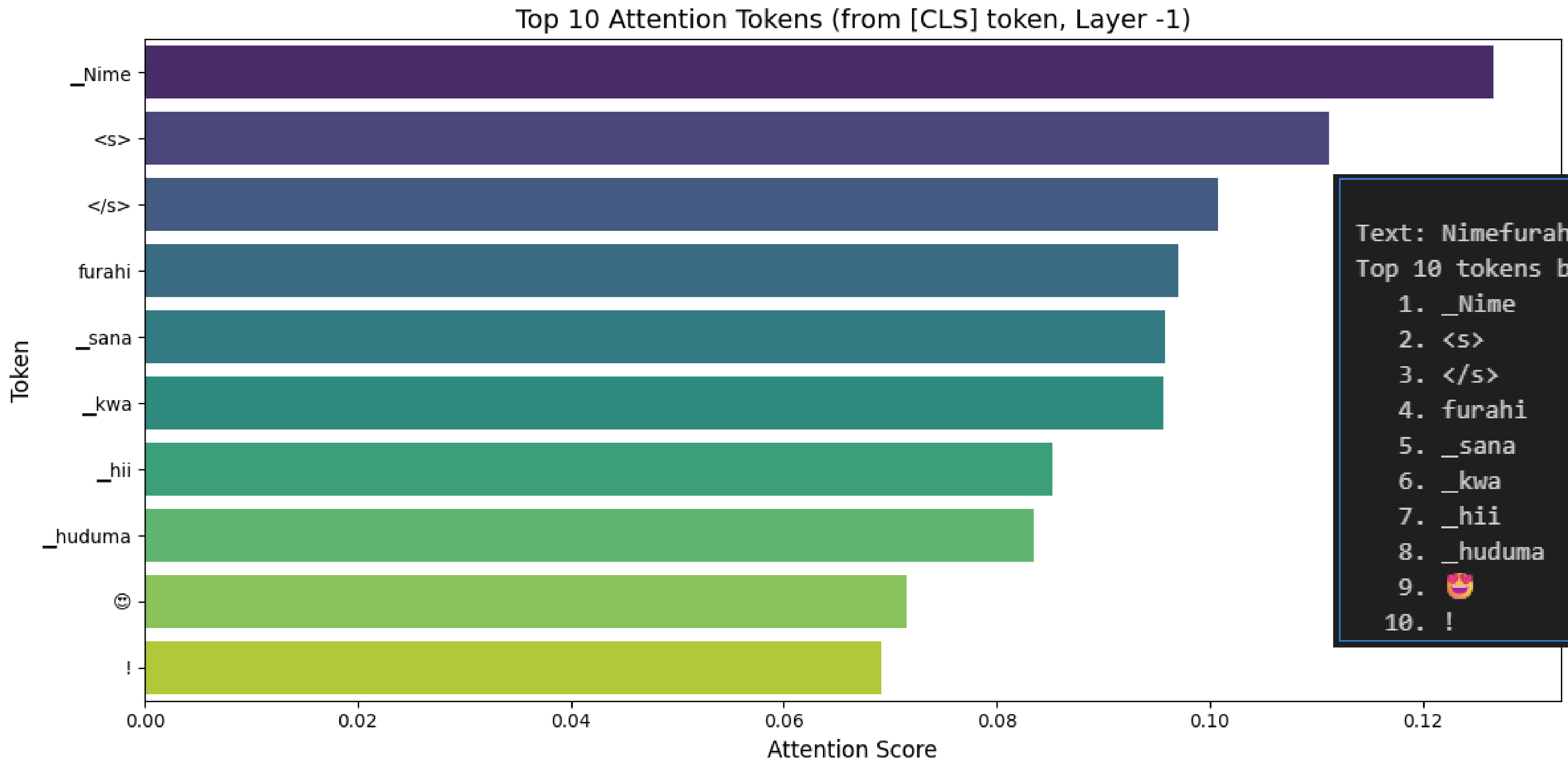
	precision	recall	f1-score	support
negative	0.65	0.64	0.64	10905
neutral	0.59	0.53	0.56	11495
positive	0.60	0.67	0.63	11907
accuracy			0.61	34307
macro avg	0.61	0.61	0.61	34307
weighted avg	0.61	0.61	0.61	34307

- Macro F1-Score: 0.6130
- Macro ROC-AUC: 0.7976

Predictions on Sample Texts do demonstrate model Behavior on Different Languages and Sentiment Classes

```
EXAMPLE PREDICTIONS - XLM-ROBERTA TRANSFORMER MODEL
____________________________________________________________
------------------------------------------------------------
Text: Nimefurahi sana kwa huduma hii! 😍
Predicted Sentiment: negative
Probabilities:
  - Negative: 0.9881
  - Neutral:  0.0101
  - Positive: 0.0017
------------------------------------------------------------

------------------------------------------------------------
Text: Service hii ni mbaya sana.
Predicted Sentiment: positive
Probabilities:
  - Negative: 0.0122
  - Neutral:  0.0201
  - Positive: 0.9677
------------------------------------------------------------

------------------------------------------------------------
Text: I am not sure how I feel about this.
Predicted Sentiment: positive
Probabilities:
  - Negative: 0.0107
  - Neutral:  0.0747
  - Positive: 0.9146
------------------------------------------------------------
```

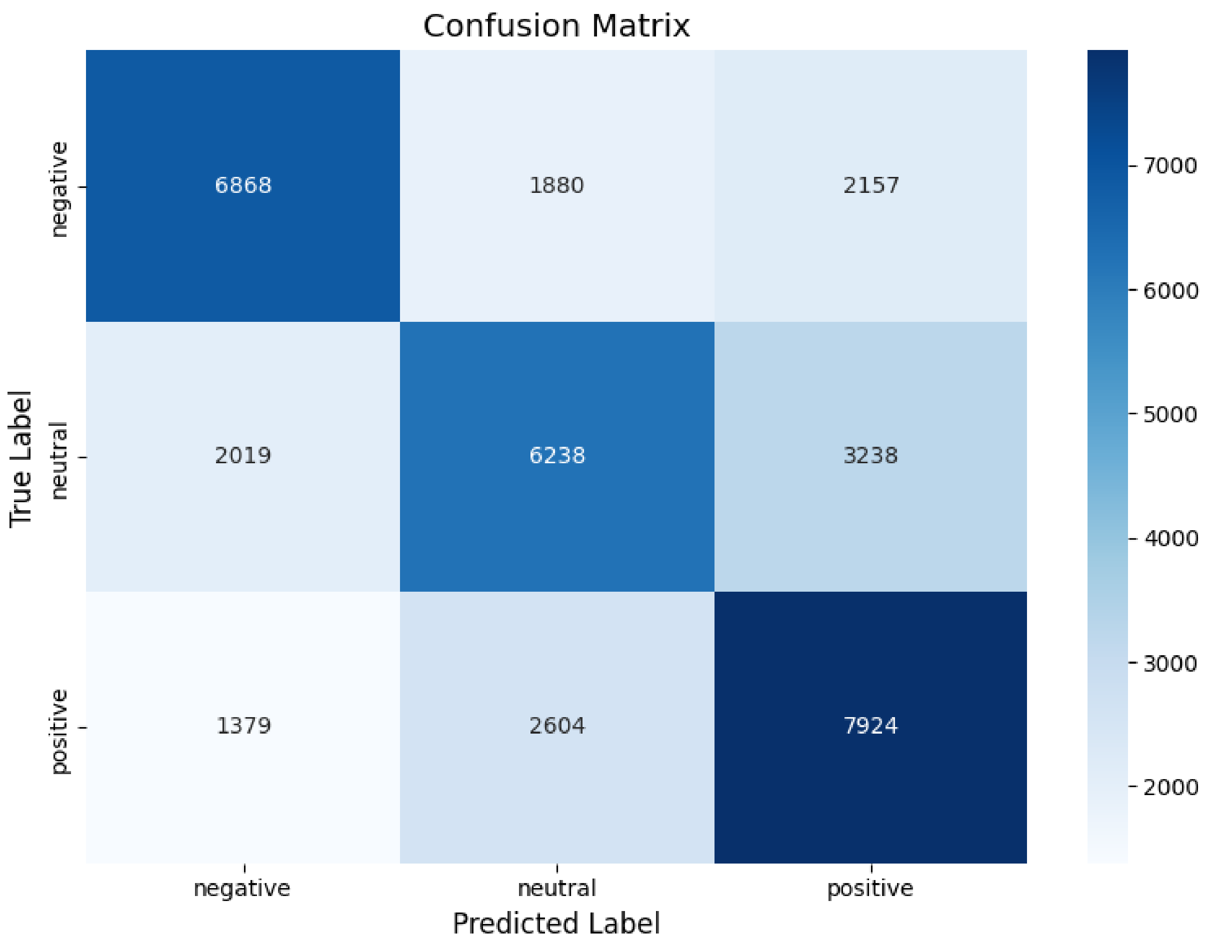
Attention Visualization XLM-ROBERTA Transformer Model




```
Text: Nimefurahi sana kwa huduma hii! 😍
Top 10 tokens by attention:
   1. _Nime                  (score: 0.1267)
   2. <s>                    (score: 0.1113)
   3. </s>                   (score: 0.1008)
   4. furahi                 (score: 0.0971)
   5. _sana                  (score: 0.0958)
   6. _kwa                   (score: 0.0957)
   7. _hii                   (score: 0.0852)
   8. _huduma                (score: 0.0834)
   9. 😍                      (score: 0.0715)
  10. !                      (score: 0.0691)
```

Ablation Studies (Batch Size, Learning Rate, Sequence Length)

```
Training model with config: batch_size=8, lr=2e-05, max_length=128
------------------------------------------------------------
[TRANS] Epoch 1/3
  Train loss: 0.8998 | Val acc: 0.6357 | Val F1: 0.6310
  New best F1: 0.6310, model saved!
[TRANS] Epoch 2/3
  Train loss: 0.7203 | Val acc: 0.6628 | Val F1: 0.6606
  New best F1: 0.6606, model saved!
[TRANS] Epoch 3/3
  Train loss: 0.5979 | Val acc: 0.6783 | Val F1: 0.6789
  New best F1: 0.6789, model saved!
Transformer training completed. Best F1: 0.6789
Total training time: 117.6 minutes (1.96 hours)
EVALUATION RESULTS
______________________________________________________________________
Accuracy: 0.6130
Macro F1-Score: 0.6129

Classification Report:
----------------------------------------------------------------------
              precision    recall  f1-score   support

    negative       0.67      0.63      0.65     10905
     neutral       0.58      0.54      0.56     11495
    positive       0.59      0.67      0.63     11907

    accuracy                           0.61     34307
   macro avg       0.62      0.61      0.61     34307
weighted avg       0.61      0.61      0.61     34307

----------------------------------------------------------------------
```



NOTE: This study did not complete due to GPU Limits

Cross-Lingual Testing: Multiple Language Pairs

NOTE:
We did not test this section due to GPU Limits

Test cross-lingual transfer with multiple language pairs:

- Train on Swahili, Test on Amharic
- Train on Swahili, Test on Pidgin English (pcm)

Testing the model's ability to transfer knowledge across languages by training on one language (Swahili) and evaluating on different target languages.

THE END

STUDENTS	AINEDEMBE DENIS +256 788-674576 dembedenisjb@gmail.com, ainedembe.denis@stud.umu.ac.ug
	MUSINGUZI BENSON +256 782 942245, musiben@gmail.com, musinguzi.benson@stud.umu.ac.ug
LECTURER	Dr. Sibitenda Harriet +256 777 056581 hsibitenda@umu.ac.ug