Sheng Guardrails Project

Investigating ways of establishing a guardrail to detect In Topic and Out of Topic discussions in a low resource *informal* language-Sheng

The context

Organization

Girl Effect provides adolescent girls and young women with essential information and links to services for immunization, mental health, nutrition, sexual and reproductive health, and early childhood development.

Context

In Kenya, Sheng, an evolving informal language that fuses Swahili and English, widely used, among the youth.

It poses significant challenges for Natural Language Processing (NLP) systems trained predominantly on standard English.

Problem

Models often have limited linguistic comprehension for low resource languages such as Sheng due to limited training data and insufficient linguistic representation.

The models are expensive to scale.

Challenges deep-dive

Challenge 1

Low resource languages

Models often have limited linguistic comprehension for low resource due to limited training data and insufficient linguistic representation

Challenge 2

Misleading or Biased Results

Without sufficient training data and insufficient linguistic representation, models might reinforce stereotypes or produce culturally insensitive outputs.

Challenge 3

Poor topic classification

Especially for cross-lingual applications arise when models attempt to perform tasks across multiple languages such as sheng

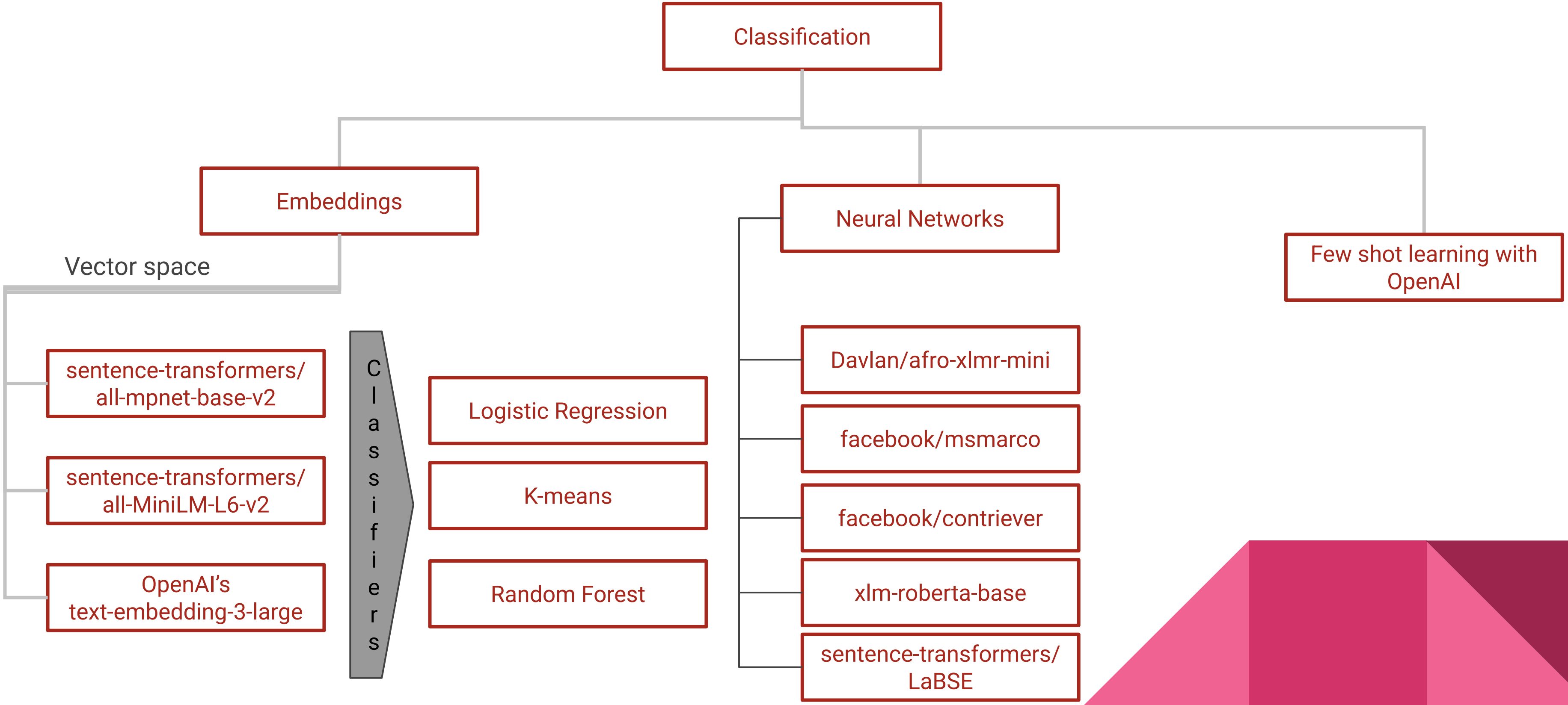
How do we classify informal languages such as sheng effectively (low cost and compute) with limited data?

Exploring Solutions

Experiments to increase classification accuracy

1. Learn shared embedding spaces for classification
2. Neural Networks
3. Few shot with OpenAI

Approach


Classification
Embeddings
Vector space
sentence-transformers/ all-mpnet-base-v2
sentence-transformers/ all-MiniLM-L6-v2
OpenAI's text-embedding-3-large
Classifiers
Logistic Regression
K-means
Random Forest
Neural Networks
Davlan/afro-xlmr-mini
facebook/msmarco
facebook/contriever
xlm-roberta-base
sentence-transformers/ LaBSE
Few shot learning with OpenAI

Classifiers

Objective: To automatically identify the decision boundary without relying on manual rule-based methods.

- Logistic regression: Supervised, and works well when the decision boundary is linear. Very unstable results
- K-Means Clustering: Unsupervised and useful in grouping embeddings without labeled data. Recommended for EDA not final classification.
- Random Forest: Able to handle complex, non-linear decision boundaries in the embedding space.

Measures

- Accuracy - Proportion of correctly predicted samples.
- F1 Score - Harmonic mean of precision and recall.A high F1 indicates the classifier does well on both classes, not just the majority.
- ROC/AUC Curve - How well the classifier separates positive vs. negative classes across all thresholds.

Understanding the data set

- *wz_qa.csv:* This is a file with 308 **sheng data** (questions and answers) that has been **vetted and labelled** already. The whole data set is IT and the questions + Answers will be combined to provide the vocabulary within the sentence embeddings. It has 2 columns, question and answer
- *wz_eng_labeled.csv:* This contains 117 rows of **data in english** with **labelled** classifications of IT/OOT.
- *human_generated_training_data.csv:* This is a 78 Row list of manually generated and **labelled sheng data**
- *wz_inference_data.csv:* This is a 25455 rows of **sheng data that is unlabelled**.
  - *human_labeled_inference_results.csv*: A subset of 450 rows of sheng data generated and manually labelled as a subset
- *small_test_data: This* is a manual generated and **labelled sheng data** listing 10 items to allow quick evaluation of accuracy through visual inspection (eyeballing).

Data usage

	Anchor Embeddings	Neural Networks	Few shot learning	Translation (Sheng-Eng)
Wz_qa.csv (308 rows)	As embeddings data ✓			✓
wz_eng_labeled.csv(117 rows)				✓
human_generated_training_data.csv(78 rows)	As embeddings data ✓	As test data ✓	As examples data ✓	
wz_inference_data.csv(24, 455)				
human_labeled_inference_results.csv (450 rows)	As train\test(0.2) data ✓	✓	✓	
Small_test_data (10 rows) as visual test	✓	✓	✓	✓

Findings and Results

Track 1: Embeddings and Cosine Similarity

- Generate embeddings using `all-mpnet-base-v2`, `all-MiniLM-L6-v2` and `text-embedding-3-large` and Classify using Logistic Regression(LR), K-means and Random Forest(RF)
- Observations:
  - In LR, the decision boundary was low at <0.25 cosine similarity, hence displayed bias towards Class 1 with low false negatives, and high false positives
  - K-means overlap zone was between 0.4-0.7, and although it seemed to have a decision boundary, the accuracy and f1 scores were poor albeit high AUC scores across `mpnet` and `minilm`. This drastically changed for `text-3-large`, where the reverse occurred.
  - In RF both classes overlap significantly between 0.1 and 0.7, making that region ambiguous. However it performed well across all sentence embeddings models

Classifier	Accuracy (mpnet / MiniLM/3-large)	F1-score (mpnet / MiniLM/3-large)	ROC - AUC (mpnet / MiniLM/3-large)
Logistic Regression	0.79/0.78/0.84	0.88/0.88/0.90	0.55/0.55/0.65
K-means	0.33/0.44/0.73	0.5/0.29/0.80	0.78/0.80/0.21
Random Forest	0.83/0.81/0.84	0.89/0.88/0.90	0.81/0.73/0.75

Track 2: Multilingual Neural Net Models

- Explored an additional NLP multilingual models with no fine tuning (used default settings).
- Observation: Despite trying several of these models the performance remained low at between 43% and 56% on the higher side.
- Conclusion:
  - Achieve a larger data set for training to test feasibility of Neural Nets for this problem set
  - Needs fine tuning of the models to achieve better results

Track 3a: OpenAI with prompt B

- Prompt B: You are a classifier that determines whether a sentence is about mental, sexual, or reproductive health. The topics include relationships, friendships, sex, abortion, menstruation, abuse, assault, health access challenges, and self-expression for young people. Classify each sentence as:
  - 1 (In Topic) if it is relevant to these areas.
  - 0 (Out of Topic) if it is unrelated.
  Here are examples: `{train_df.to_dict(orient='records')}`
  Sentence to classify: `{item}`
  Answer:

	Accuracy	F1 Score	AUC Score
gpt-4o-mini	0.833	0.835	0.86
gpt-3.5-turbo-instruct	0.7430	0.7436	0.73

Track 4: Translations

- Translated the smaller data set to English using OpenAI, used text-3-large embeddings and used cosine similarity with Logistic Regression
- **Observation:** Despite the transformation, accuracy and F1 remained below 50%, suggesting translation introduced semantic drift(change or distortion in the meaning of a word, phrase, or sentence when it's transformed) without improving decision boundary clarity.

Observations and Conclusions

Observations

- Semantic meanings in Sheng across both IT and OOT classes are close to each other that it limits the ability of the classification
- `all-mpnet-base-v2` performed close to `text-embedding-3-large` has high F1 scores because it is fine-tuned on semantic textual similarity tasks so its embeddings naturally group semantically similar texts closer in vector space.
- LR and RF performed above 70% with `all-mpnet-base-v2` sentence embeddings while K-means struggled
- Few shot with OpenAI's gpt-3.5-turbo-instruct which was added as a benchmark provided similar results to LR and RF
- Worth noting is OpenAI's gpt-4o-mini showed a higher results above 80% in both accuracy and F1 scores
- Random forest with AUC of >70% and accuracy/f1 scores >81% implies a good useable model, while K-means and LR need more improvement
- Test set was randomly generated, however it maintained 19 Class 0(Out of Topic) and 71 (In topic) records which represent a class imbalance. This led to under-training the model on the minority class (leading to high false negatives)

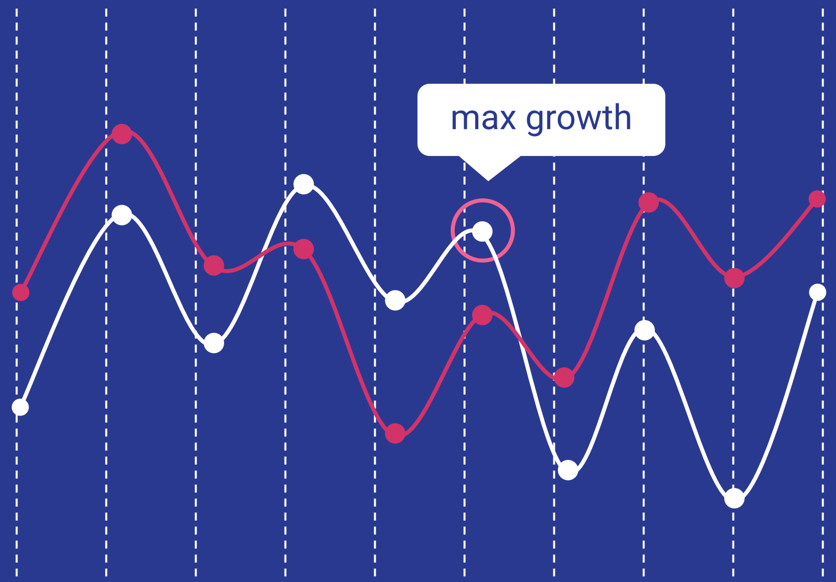
Conclusion

- With a curated training data set, the traditional classification models (Logistic Regression, and Random Forest) perform well for topic classification in a low resource language such as Sheng similar to gpt-3.5-turbo-instruct
- The models should be trained using a combination of english, swahili and sheng datasets for improved results

Further research

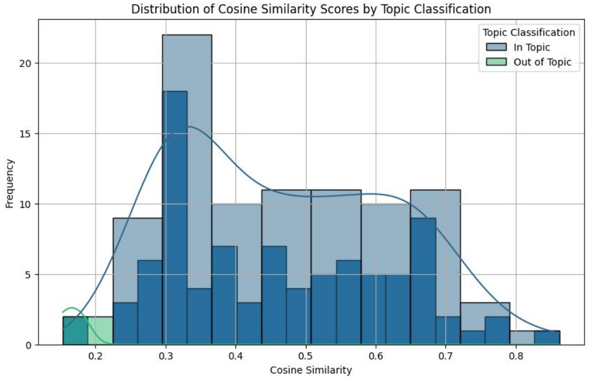
- Use of logistic regression with regularization to improve model performance
- K-means with k=3 to introduce a grey zone where human intervention may be required to determine if IT and OOT or an alternative model applied
- OpenAI with prompt B was used as opposed to the regular prompt A due to tokenization rate limits on inputs. Recommendation is to handle token sizes in the few shot approach

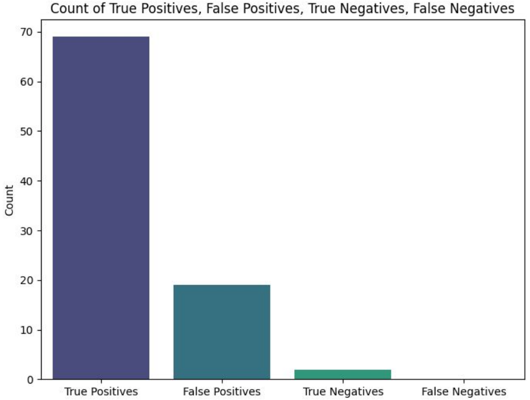
APPENDIX



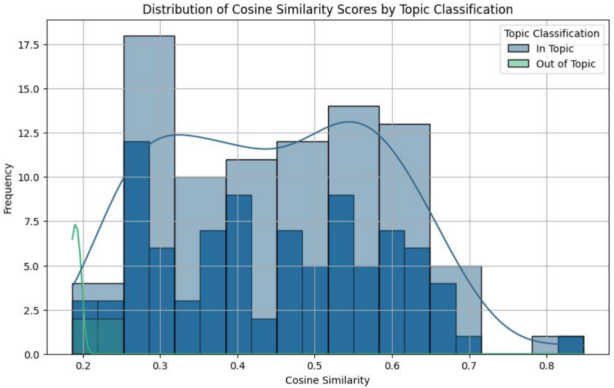
Logistic Regression Distribution

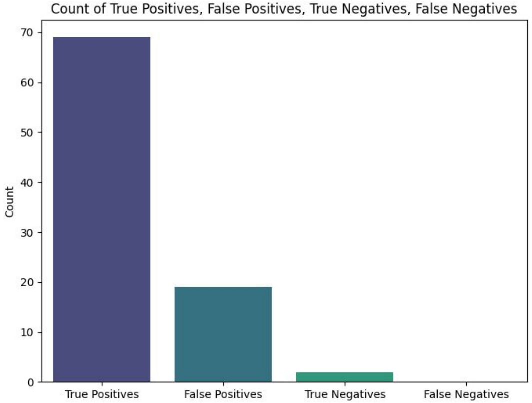
all-mpnet-base-v2



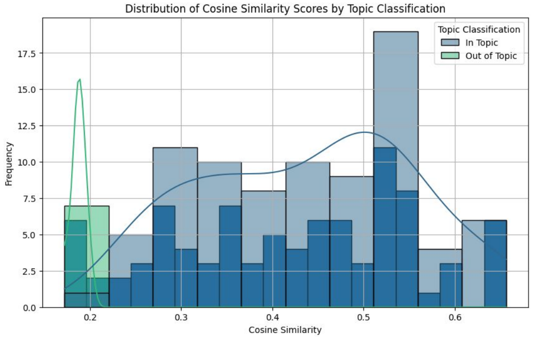


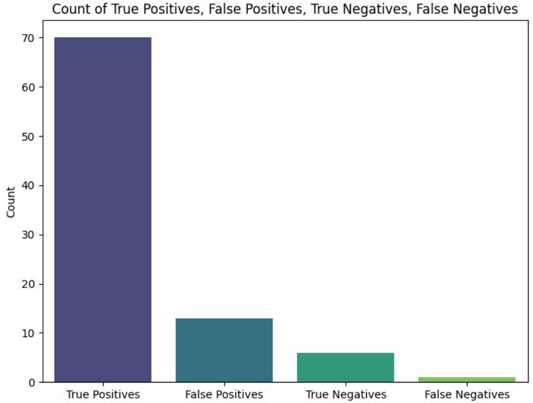
all-MiniLM-L6-v2





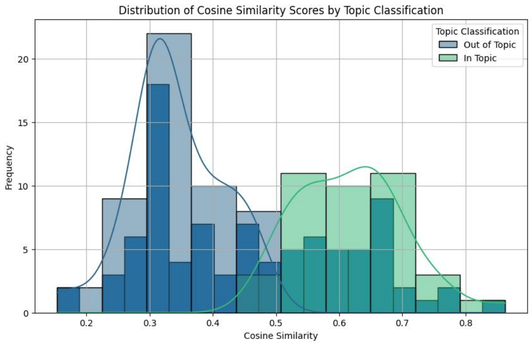
text-embedding-3-large

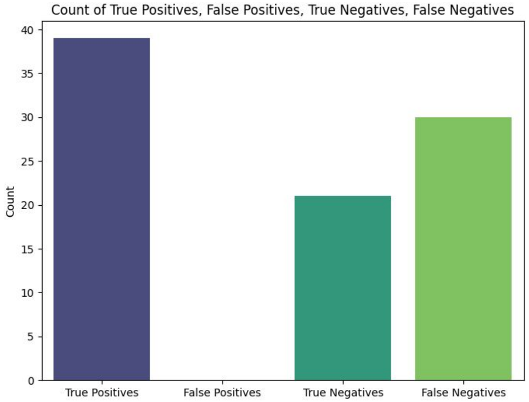




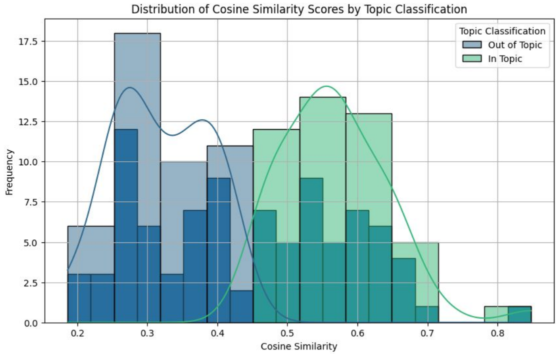
K-means Distribution

all-mpnet-base-v2

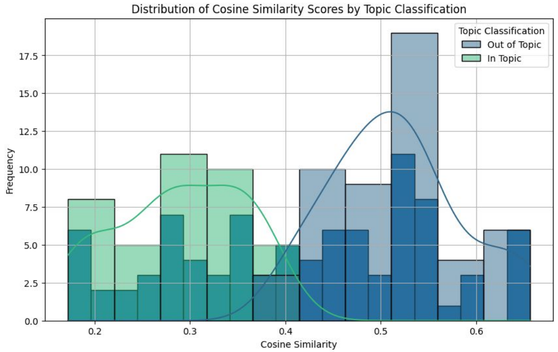


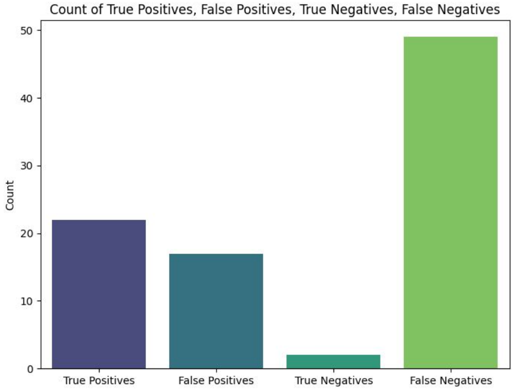


all-MiniLM-L6-v2



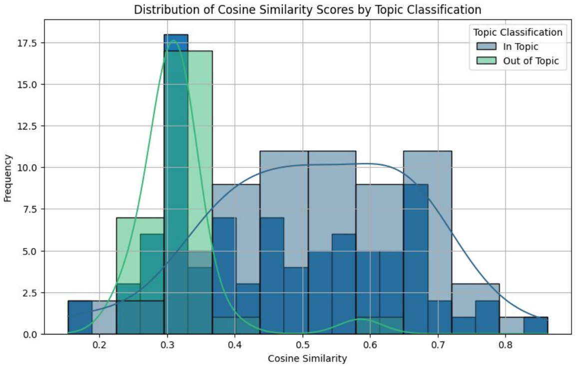
text-embedding-3-large

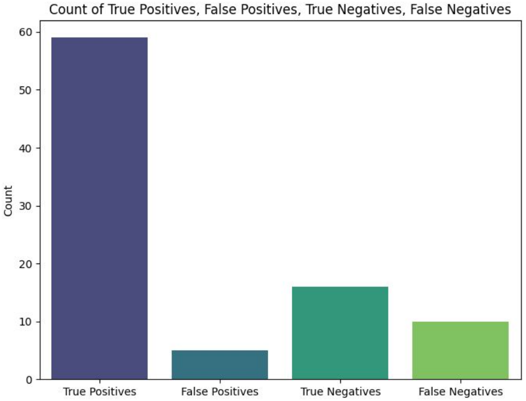




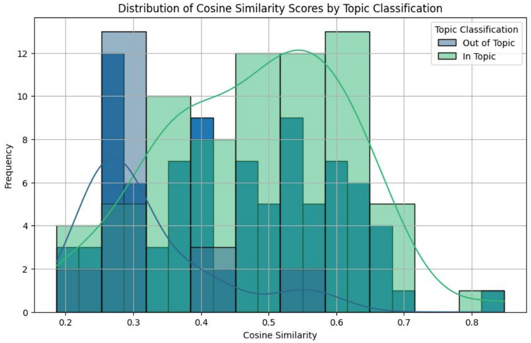
Random Forest Distribution

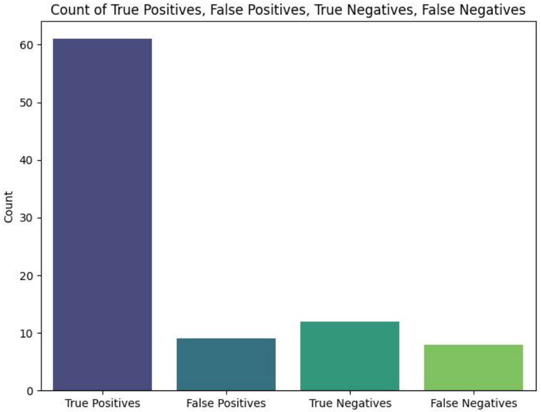
all-mpnet-base-v2



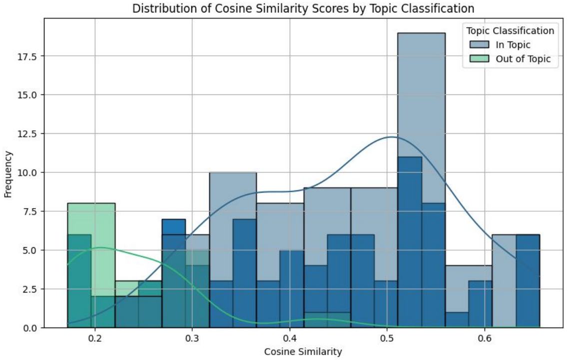


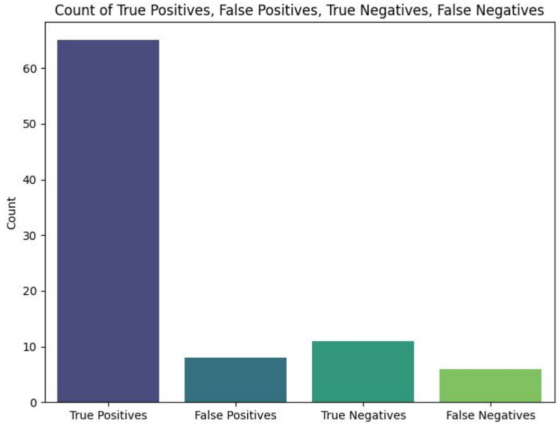
all-MiniLM-L6-v2





text-embedding-3-large





Track 2: Multilingual Neural Net Models

Track 4: Explored an additional NLP multilingual models with no fine tuning (used default settings).

Observation: Despite trying several of these models the performance remained low at between 43% and 56% on the higher side.

Conclusion:

- Achieve a larger data set for training to test feasibility of Neural Nets for this problem set
- Needs fine tuning of the models

Model	Validation loss	Accuracy	F1 score
Davlan/afro-xlmr-mini	0.740749	0.500000	0.333333
facebook/msmarco	1.381521	0.500000	0.333333
facebook/contriever	0.651121	0.562500	0.458937
xlm-roberta-base	0.707880	0.500000	0.333333
sentence-transformers/LaBSE	0.959065	0.500000	0.333333