RideKE: Leveraging Low-Resource, User-Generated Twitter Content for Sentiment and Emotion Detection in Kenyaǹ Code-Switched Dataset

Naome A. Etori and Maria L. Gini
Department of Computer Science and Engineering
University of Minnesota -Twin Cities
{etori001, gini} @umn.edu

Abstract

Social media has become a crucial open-access platform for individuals to express opinions and share experiences. However, leveraging low-resource language data from Twitter is challenging due to scarce, poor-quality content and the major variations in language use, such as slang and code-switching. Identifying tweets in these languages can be difficult as Twitter primarily supports high-resource languages. We analyze Kenyan code-switched data and evaluate four state-of-the-art (SOTA) transformer-based pretrained models for sentiment and emotion classification, using supervised and semi-supervised methods. We detail the methodology behind data collection and annotation, and the challenges encountered during the data curation phase. Our results show that XLM-R outperforms other models; for sentiment analysis, XLM-R supervised model achieves the highest accuracy (69.2%) and F1 score (66.1%), XLM-R semi-supervised (67.2% accuracy, 64.1% F1 score). In emotion analysis, DistilBERT supervised leads in accuracy (59.8%) and F1 score (31%), mBERT semi-supervised (accuracy (59% and F1 score 26.5%). AfriBERTa models show the lowest accuracy and F1 scores. All models tend to predict neutral sentiment, with Afri-BERT showing the highest bias and unique sensitivity to empathy emotion. [1]

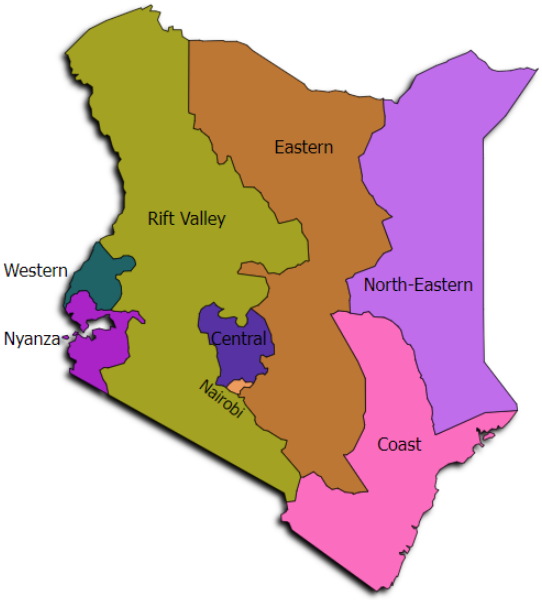


Figure 1: **Geographical representation of RideKE:** *diverse local accents collected in tweets, such as Rift Valley (e.g., Eldoret, Nakuru), Central (e.g., Nyeri, Kiambu), Nairobi (e.g., Kasarani, Kileleshwa), Western (e.g., Kakamega, Bungoma), Nyanza (e.g., Kisumu, Kisii), Eastern (e.g., Machakos, Embu) Coast (e.g., Mombasa, Malindi), and North-Eastern (e.g., Garissa, Mandera).*

1 Introduction

Kenya, reflecting Africa's extensive multilingual diversity, offers a unique insight into the continent's rich linguistic heritage, standing as a focal point of language contact, expansion, and diversity. It is home to many languages that bridge its vibrant storytelling, poetry, song, and literature and exemplifies Africa's linguistic wealth, albeit on a more localized scale. With over 40 languages grouped into Bantu, Nilotic, and Cushitic, Kenya's linguistic landscape is diverse and dynamic (Dwivedi, 2014; Carter-Black, 2007; Banks-Wallace, 2002).

Central to linguistic diversity is the co-official language status of English and Kiswahili, with the latter spoken by the majority and enjoying near-equal prominence with English. However, the linguistic equilibrium faces challenges from Sheng, a language that blends English, Kiswahili, and words from other ethnic languages that initially were used in Nairobi Eastlands slums. Sheng emerged as a sociolect among urban youth in the city's working-class neighborhoods and has since spread across various social and age groups. Hence, it is an inte-

[1] https://github.com/NEtori21/Ride_hailing_project

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 234–249

Tweets	Sentiment	Emotion
Uber kenya did your App stop accepting cards for package deliveries? I have had two riders this morning cancel picking a package because they want me to pay cash.	Negative	Frustration
Thank you for the love and support and for the feedback as well. Tell all your friends to ride a littleCab. Buy Kenyan, build Kenya.	Positive	Love
Uber drivers are not employees of Uber Kenya Uber is only an app. The link between you as a rider and the driver. But yes they should look after them because the drivers keep them afloat.	Neutral	Neutral
A ride will be canceled for one reason or another and both parties should have the liberty to. Sometimes clients will cancel due to the proximity of the driver and other times because the driver is unreachable.	Neutral	Neutral
Hope everyone making the most of this awesome Uber kenya Jan offer! Spread the word! Loving it. #Uber kenya	Positive	Happy
Giving drivers right to refer the rider to another driver then that is totally not a good idea. Some drivers are connecting while he like really far from you, he wastes time, then after more than 5 mins refers another driver	Neutral	Happy
Greater experience for Uber riders with new product	Positive	Happy
I am reporting your driver for taking payment twice. I had ordered an Uber for a friend with payment with a card and then he tells the passenger to pay via Mpesa.	Negative	Frustration
I also stopped using Uber kenya after I was charged for cancelling a trip as per the drivers request. Little cab iko tu sawa.	Negative	Frustration
Crooked policies. Uber kenya. I think you need to sort out your service.	Negative	Angry
Honestly, Am disappointed with them. kucancel trips ndio wanajua lately.	Negative	Frustration

Table 1: Sample Tweets with Sentiment and Emotion Labels.

gral part of Kenyan culture, influencing the traditional dominance of English and Kiswahili (Barasa, 2016; Momanyi, 2009; Mazrui, 1995).

In recent years, language diversity has also been mirrored in the urban transportation sector, primarily due to the growth of Ride-Hailing Services (RHS) such as Uber, Bolt, and Little Cab. These services have rapidly transformed from urban novelties to essential components of daily mobility for many Kenyans, connecting remote areas with vibrant urban cities. However, with the entry of global giants like Uber in 2015, followed by Bolt and the local contender Little Cab, this transformation is not just physical; it extends into digital and social media platforms such as Twitter.

Since many languages are spoken across Kenya, each population has its own dialect. Hence, code-switching is common in these new forms of communication, where speakers alternate between two or more languages in one conversation (Kanana Erastus and Kebeya, 2018; Santy et al., 2021; Angel et al., 2020; Thara and Poornachandran, 2018). Analyzing sentiment and emotions in code-switched language context is critical in the broad natural language processing (NLP) field, for example, creating systems that can predict emotional states from text to speech which can be applied in various use cases, such as measuring consumer satisfaction (Ren and Quan, 2012), natural disasters (Vo and Collier, 2013), marketing strategy (Zamani et al., 2016), e-learning (Ortigosa et al., 2014), e-commerce(Jabbar et al., 2019) and psychological states (Aytuğ, 2018). However, despite this linguistic richness, African languages remain significantly underrepresented in NLP research (Muhammad et al., 2023a). Although NLP research has made extensive progress and demonstrated broad utility over the past two decades, the focus on African languages has been limited. This disparity is often attributed to the scarcity of high-quality, annotated datasets for these languages.

Recently, researchers (Muhammad et al., 2023a)[2] have focused on addressing this challenge by introducing a comprehensive benchmark with over 110,000 tweets across 14 African languages, Swahili among them, and introduced the first Africentric SemEval Shared task (Muhammad et al., 2023b). Various studies have evaluated the performance of state-of-the-art (SOTA) transformer models on African languages, highlighting unique challenges and opportunities (Aryal et al., 2023).

However, research on social media NLP analysis for RHS datasets mainly targets high-resource languages. NLP for low-resource languages is constrained by factors like NLP research's geographical and language diversity (Joshi et al., 2020). Using pre-trained transformer models, we introduce RideKE, a sentiment and emotion analysis dataset for African-accented English code switched with Swahili and Sheng.

[2] https://github.com/afrisenti-semeval/afrisent-semeval-2023

Code-switched Reference	English Translation
I recently interacted with one Uber driver who told me that **huko ni mbali, lazima uongeze pesa**. Different from the estimate on the app. He almost dropped me midway because I argued that it wasn't fair. **Hawa madere ni wazimu walai.**	I recently interacted with one Uber driver who told me that the place is far, you have to add money. Different from the estimate on the app. He almost dropped me midway because I argued that it wasn't fair. These drivers are crazy, really.
In Mombasa, they ask you how much the App has displayed as the cost, then tell you it's too low, madam **unaona utaongeza ngapi**, **hiyo pesa ni kidogo**	In Mombasa they ask you how much the app has displayed as the cost then tell you it's too low, madam how much extra?, That's little money

Table 2: Example of code-switched sentences in Tweets

Our dataset contains over 29,000 tweets, each sentiment classified as either positive, negative, or neutral, and emotions classified as frustration, happy, angry, sad, empathy, fear, love, and surprise. The dataset represents one location, Kenya, as shown in Table 1. Our goal is to advance research in low-resource languages.

The experiments in this paper are designed to allow us to answer the following specific questions:

1. How do pretrained language models enhance the detection and representation of Kenyan low-resource languages and accents in modern NLP tools?
2. How does the performance of sentiment and emotion detection varies across different pretrained transformer-based models?
3. How effective are different transformer-based models in performing sentiment and emotion detection on the low-resource (RideKE) dataset using semi-supervised learning?

Our paper makes the following contributions as we address these questions:

- We use semi-supervised learning to classify sentiments and emotions. We compare four SOTA transformer-based models and provide a detailed model performance analysis.
- We contribute a partially curated human-annotated labeled public dataset with over 29,000 tweets from the RHS domain. This is Kenya's first-ever code-switched sentiment and emotion dataset in the RHS domain. It contributes resources to low-resource areas, which can be used for other analyses.

2 Literature Review

2.1 Sentiment Analysis on Social Media

Sentiment analysis (SA) emerged as a significant field early in the 2000s (Das and Chen, 2001; Nasukawa and Yi, 2003). SA (Dave et al., 2003; Pang et al., 2008) aims to determine the attitudes, opinions, or emotions expressed in text on specific topics or entities (Liu, 2022) and has become an increasingly popular research area. Due to higher user-generated content available on social media, understanding sentiment in text cannot be overstated (Naseem and Musial, 2019).

Diverse strategies to accurately interpret and classify user sentiments have been employed. For example, lexicon-based approaches, like SENTIWORDNET (Baccianella et al., 2010) and AFINN (Nielsen, 2011), used predefined word lists to classify text sentiment. While effective in some applications, these methods often struggled with context and nuance. Rule-based systems (Suttles and Ide, 2013) further enhanced this method by applying contextual rules to detect sentiment nuances, including handling negations (Taboada et al., 2011).

Advancements in Machine learning (ML) (Pang et al., 2002), such as supervised techniques trained on large amounts of labeled sentiment datasets, offer another powerful avenue for SA. Hence, the exploration of semi-supervised methods in SA could leverage unlabelled data to address the challenge of data annotation and labeling (Vo and Zhang, 2015; Hwang and Lee, 2021). Deep learning approaches such as Convolutional Neural Networks (CNN) (Chen, 2015) have significantly advanced SA capabilities. However, SA on social media poses unique challenges compared to more traditional domains due to the informal and conversational nature of the text (Medhat et al., 2014; Naseem and Musial, 2019).

2.2 Code-Switching on Low-resource

Code-switching, the practice of alternating between two or more languages or dialects within a conversation, is particularly prevalent in multilingual communities and has become increasingly visible

on social media platforms (Poplack, 2000; Scotton, 1993; Danet and Herring, 2007). It presents unique challenges and opportunities for NLP (Barman et al., 2014). Most NLP research traditionally focuses on high-resource languages like English, leaving low-resource languages underrepresented (Strassel and Tracey, 2016; Adelani et al., 2021). This gap is more pronounced in African and code-switched languages due to linguistic variability (Adelani et al., 2021). Therefore, high-resource language techniques may underperform on low-resource language data (Lewis, 2014). The study in (Lee and Wang, 2015) emphasizes the importance of analyzing emotions in code-switching data. The use of Generative Pre-trained Transformers (GPT) to generate synthetic code-switched data has been proposed to address data scarcity (Terblanche et al., 2024). A recent survey (Winata et al., 2022) revealed that until October 2022, only a few papers from the ACL Anthology and ISCA Proceedings focused on code-switching research in African languages. For South African languages (Niesler et al., 2018; Niesler and De Wet, 2008) the first dataset was presented in 2018. Even though Swahili-English code-switching has been studied in a few papers (Piergallini et al., 2016; Otundo and Grice, 2022), no datasets are available.

2.3 Transformer-based Pretrained Models

Transformer-based architectures (Vaswani et al., 2017), such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), have gained popularity owing to their effectiveness in learning general representations using large unlabelled datasets (Matthew, 2018) that can further be fine-tuned for downstream tasks (Gururangan et al., 2020; Bhattacharjee et al., 2020). Hence, it has become the foundation for many NLP tasks (Bhattacharjee et al., 2020).

Pretrained language models are trained on large, diverse datasets (Raffel et al., 2020). For example, RoBERTa (Liu et al., 2019) was pretrained on over 160GB of uncompressed text, from BOOKCORPUS (Zhu et al., 2015) and CommonCrawl English dataset (Nagel, 2018). These models learn representations that perform well across various tasks, handling datasets of different sizes from diverse sources while remaining easily understandable (Wang et al., 2019). Examples of a few applications in low-resource include improving speech recognition accuracy (ASR) (Olatunji et al., 2023), machine translation (MT) (Wang et al., 2024) and SA (Muhammad et al., 2023a).

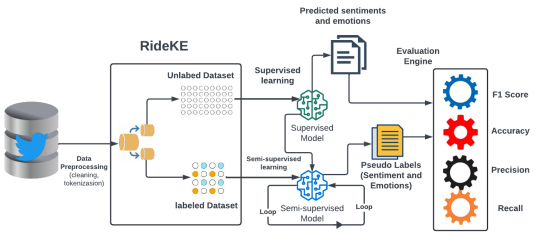


Figure 2: **Methodology:** *Overview of the RideKE sentiment and emotion analysis framework. Unlabeled and labeled datasets are preprocessed and used to train supervised and semi-supervised models for sentiment and emotion prediction. The semi-supervised learning loop generates pseudo labels for evaluation of performance.*

3 Methods and Datasets

3.1 Overview of RideKE Dataset

RideKE dataset. as shown in Table 1 and 2, includes a blend of Kenyan-accented English, approx. (70%), with a minority mix of Swahili and Sheng (30%). The dataset includes a total of 29,623 entries across 12 distinct columns. See Table 13 in the Appendix.

3.2 Data Collection

We used a systematic scraping process using the snscrape python library [3] which allows for querying and retrieving tweets based on specified criteria. We targeted three keyword search terms—#UBER-Kenya, #BOLT-kenya, and #LITTLECAB, from January 2017 to April 2023, capturing not only the tweet texts but also other essential metadata such as user engagement metrics (likes, retweets, replies), user account details (followers, following, tweet counts), and relational markers (hashtags, user mentions). Initially, the data was in a dictionary format but it was later converted to DataFrame using pandas and preserved in a CSV format to ensure reproducibility.

3.2.1 Geo-based data collection

The tweet's location metadata was crucial in determining the regional focus of our study. We referenced Kenya's location as shown in Table 3. To ensure uniformity, we used a simple yet effective keyword filtering normalization technique to address location inconsistencies as shown by the diverse representations of Nairobi in the dataset

[3] https://pypi.org/project/snscrape/1

shown in Table 3. To isolate the relevant tweets, we applied a filter on the `user_location` field to include only locations mentioning `Kenya` and discard entries with missing data and all those with no location. We assessed the frequency distribution of different locations using value count function.

Location	Tweet Count
Kenya	18974
Nairobi, Kenya	11960
Not specified	10868
Nairobi	4776
Nairobi, Kenya	620
nairobery	1
Africa, Nairobi Kenya	1
Mt. Meru	1
3rd Parklands	1
New Jersey	1

Table 3: **Tweet Counts by location:** *We only included locations mentioning Kenya*

3.3 Language Detection

We used `langdetect` [4] Python library to detect languages within text. It revealed diverse languages, `English` being the most prevalent, then `Indonesian`, `Swahili` and others as shown in Table 10. For the Sheng language, native speakers manually detected the language. We only kept English (code-switched) for our analysis.

3.4 Data Preprocessing

Tweets often feature slang, abbreviations, and non-alphanumeric characters such as hashtags and emojis, contributing to the data's unstructured nature (Adebara and Abdul-Mageed, 2022). We implemented a refined text preprocessing pipeline to enhance data consistency and accurate analysis. The pipeline standardizes data by converting text to strings, trimming whitespace, lowering case, and expanding contractions to preserve semantic integrity. The text is then normalized by reducing repeated characters, removing punctuation, newlines, and tabs, and then tokenizing.

3.5 Data Annotation

Inspired by (Raffel et al., 2020) established guidelines, we created a set of annotation guidelines for emotion annotations to ensure a standardized and high-quality approach in our labeling efforts, as shown in Table 12. We added a 'frustration' label and used 'happy' instead of 'joy.' For the sentiment annotation, we adhered to the established annotation framework detailed by (Mohammad, 2016). However, human annotation is time-consuming and costly. We employed two Kenyan volunteer annotators fluent in English, Swahili, and Sheng. One holds a bachelor's degree in political science and the other in computer science. They received a small token of appreciation for their efforts. We ensured the annotator's comprehension of the task. Two annotators labeled the same dataset entries to enhance quality. Each labeled 1,554 tweets with sentiment labels (positive, negative, neutral) and emotion labels (sadness, happy, love, anger, fear, surprise, frustration, and neutral).

3.5.1 Annotation Quality Control

We used Cohen's Kappa (Artstein, 2017)[5] as our primary metric for assessing the level of inter-annotator agreement between the two annotators. It is perfect for categorical items, such as sentiment and emotion labels. Cohen's Kappa provides a means to compute an inter-rater agreement score that accounts for the probability of random agreement:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the observed agreement, and P_e is the expected agreement by chance.

To assign the final sentiment and emotion label to each tweet, we employed a majority voting method (Davani et al., 2022) to determine the final label of the tweet (Mohammad, 2022). Instances of complete disagreement among annotators were resolved by involving a lead annotator and applying a majority rule rather than omitting them from the dataset. We found a Cohen's Kappa coefficient of 0.60 for sentiment classification tasks. Cohen's Kappa score for the emotion annotations is approximately 0.67, which indicates a substantial level of agreement beyond chance and suggests a good degree of consistency in their annotations.

3.5.2 Data Splits

The dataset was split into three sets (A, B, and C) as shown in the dataset division Table 4. We used ChatGPT (Brown et al., 2020) for automatic labeling to augment the training dataset and increase training labels since we had only two human annotators. Set A provided Ground truth labels for initial supervised training. Set B is the test dataset

[4] https://pypi.org/project/langdetect/1

[5] https://github.com/zyocum/cohens_kappa

that is manually annotated by human annotators. Set C represented the unlabelled dataset Used in a semi-supervised training loop, with empty rows and duplicates removed, labels standardized and encoded.

Set	Description	Details
Set A	553 human, 636 ChatGPT	Supervised Train
Set B	2,000 human	Testing
Set C	27,090 unlabelled	Semi-supervised

Table 4: Dataset Division

3.6 Semi-supervised Learning Phase

Semi-supervised learning (SSL) offers a framework for utilizing large amounts of unlabelled data when obtaining labels is expensive (Chapelle et al., 2006; Learning, 2006) as applied to our case. Research shows SSL improves performance on different machine learning tasks such as text classification and machine translation (Najafi et al., 2019). SSL connects supervised and unsupervised learning by utilizing a small fraction of labelled data alongside a larger pool of unlabeled data to improve learning accuracy. SSL has been widely studied to show effectiveness for a wide range of low-resource applications, such as in text-to-speech synthesis (TTS) (Saeki et al., 2023), speech recognition (Du et al., 2023; Thomas et al., 2013), machine translation(Pham et al., 2023; Singh and Singh, 2022), POS-Taggers (Garrette et al., 2013), and sentiment classification (Gupta et al., 2018). Our work extends the application of SSL to sentiment and emotion classification tasks. We seek to mitigate this limitation by leveraging labeled and unlabeled data to train pretrained models. We used accuracy, precision, recall, and F1 scores to evaluate the models' performance.

4 Experiments

4.1 Models and Architecture

We evaluate four transformer-based models in our experiments: **DistilBERT** (Sanh et al., 2019), a smaller and faster version of BERT; **mBERT** (Devlin et al., 2018), a multilingual version of BERT trained on 104 languages; **XLM-RoBERTa** (Conneau et al., 2019), a multilingual model trained on 100 languages with improved performance; and **AfriBERTa large** (Ogueji et al., 2021), a model specifically designed for African languages to address the unique linguistic challenges in this region. Each model was trained on supervised and semi-supervised learning on sentiment and emotion classification tasks. The initial supervised training and subsequent semi-supervised fine-tuning were conducted separately for each model.

4.2 Experimental Setup

4.2.1 Supervised Learning Phase

In supervised training, we utilized the human-annotated, well-curated labeled dataset. We used batches ranging from 16 to 64 depending on the model sizes, optimizing for computational efficiency. A combined categorical cross-entropy loss shown in Figure 3 function, with equal weighting for sentiment and emotion tasks, guided the model toward effective multitasking. We applied a dropout rate of 0.1 for each model to prevent overfitting and enhance generalization. We employed the Adam optimizer, with a learning rate $1e-5$ through 10 epochs of training and monitoring. Initially, the four transformer-based models were fine-tuned on a dataset with 1,189 labeled tweets. We then evaluated the model.

4.2.2 Semi-supervised Learning Phase

Our goal in using SSL is to leverage the vast, unlabeled datasets to mitigate the high cost of human annotations. Following an initial supervised learning phase, each transformer-based model underwent a semi-supervised training loop. In this loop, the models dynamically labeled the unlabeled dataset based on their predictions, generating a pseudo-labeled dataset. We employed a dynamic threshold, set at the 75th percentile of the models' probability predictions across all classes for each batch, to ensure only high-confidence predictions were used for labeling. Samples with predictions below this threshold were excluded to minimize the inclusion of erroneous labels in the training data.

We extended the semi-supervised training loop over 4 epochs, a duration we empirically selected to refine the models' generalization capabilities without causing performance degradation due to overtraining, as indicated by either worsening or plateauing loss. We carefully chose the hyperparameters to ensure optimal training dynamics and model performance.

We set the learning rate at 1e-5 and dynamically adjusted it using a learning rate scheduler during training to optimize generalization and reduce overfitting. The batch size varied between 16 and 64, depending on the specific transformer model, to en-

Model	Sentiment				Emotions			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
DistilBERT supervised	0.578	0.598	0.629	0.546	0.598	0.334	0.315	0.310
DistilBERT semi-supervised	0.553	0.585	0.598	0.516	0.544	0.264	0.266	0.252
mBERT supervised	0.638	0.621	0.663	0.596	0.592	0.253	0.298	0.265
mBERT semi-supervised	0.635	0.622	0.661	0.598	0.594	0.297	0.317	0.297
XLM-R supervised	0.692	0.665	0.723	0.661	0.658	0.343	0.267	0.258
XLM-R semi-supervised	0.672	0.644	0.702	0.641	0.620	0.334	0.248	0.230
AfriBERTa large supervised	0.398	0.500	0.479	0.358	0.604	0.163	0.191	0.157
AfriBERTa semi-supervised	0.413	0.534	0.491	0.366	0.556	0.145	0.177	0.142

Table 5: **Model Performance Evaluation on Sentiment and Emotion Analysis Tasks.** *Performance evaluation of supervised and semi-supervised training for sentiment and emotion analysis across models. Results represent averages over multiple runs.*

Model	Negative			Neutral			Positive		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
DistilBERT supervised	0.920	0.385	0.543	0.284	0.635	0.392	-	-	-
DistilBERT semi-supervised	0.901	0.325	0.478	0.268	0.604	0.371	-	-	-
mBERT supervised	0.906	0.467	0.616	0.330	0.587	0.423	-	-	-
mBERT semi-supervised	0.873	0.443	0.588	0.363	0.628	0.460	-	-	-
XLM-R supervised	0.921	0.563	0.699	0.417	0.714	0.526	-	-	-
XLM-R semi-supervised	0.850	0.524	0.648	0.392	0.712	0.506	0.691	0.871	0.771
AfriBERTa large supervised	0.794	0.100	0.178	0.144	0.492	0.223	-	-	-
AfriBERTa semi-supervised	0.874	0.096	0.174	0.171	0.560	0.261	0.558	0.817	0.663

Table 6: **Model Performance Evaluation on Sentiment classification Tasks Labels.** *Performance evaluation for Negative, Neutral, and Positive sentiments across various models. A dash (-) indicates missing values, i.e., the models did not predict all positive sentiment instances. The results represent averages over multiple runs.*

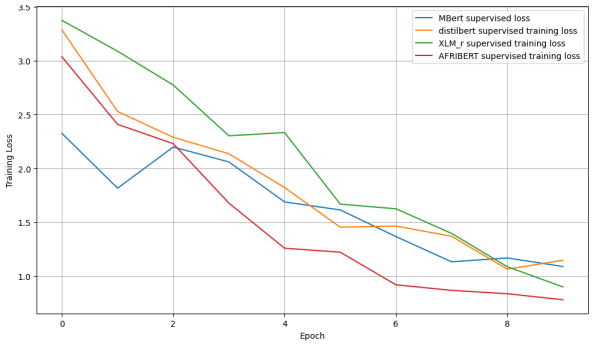
sure computational efficiency. We used a combined loss function shown in Figure 3 for sentiment and emotion analysis and applied a dropout rate of 0.1 to prevent overfitting. We employed the Adam optimizer with a learning rate of 1e-5 and no weight decay.

5 Results and Discussions

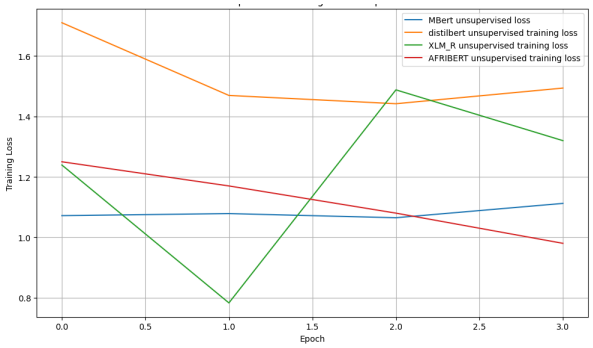
5.1 Sentiment Analysis

Table 5 summarizes the performance of all models on sentiment analysis. XLM-R supervised achieves the highest overall performance with an accuracy of 62.5% and an F1-score of 66.7%. This is followed closely with semi-supervised XLM-R, which has an accuracy of 62.1% and an F1-score of 68.3%. However, DistilBERT supervised performance falls behind with an accuracy of 57.8% and an F1-score of 54.6%. On the other hand, mBERT models show consistency between supervised and semi-supervised training, maintaining average F1-scores of 59.8% and 59.6%, respectively. AfriBERTa models struggled, with the supervised learning achieving an F1-score of 35.8%, and overall poorest performance across all metrics.

The detailed performance metrics for negative, neutral, and positive sentiment classification are



(a) Supervised Loss



(b) Semi-supervised Loss

Figure 3: **Training loss** *(a) supervised and (b) semi-supervised learning.*

Metrics	Neutral			Frustration			Happy		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Distilbert_supervised	0.130	0.176	0.150	0.444	0.364	0.400	0.000	0.000	0.000
Distilbert_semi_supervised	0.141	0.121	0.131	0.132	0.227	0.167	0.000	0.000	0.000
mBERT_supervised	0.043	0.059	0.050	0.000	0.000	0.000	0.000	0.000	0.000
mBERT_semi_supervised	0.284	0.234	0.256	0.100	0.045	0.063	0.000	0.000	0.000
XLM_R_supervised_training	1.000	0.118	0.211	0.000	0.000	0.000	0.000	0.000	0.000
XML_R_semi_supervised	0.571	0.037	0.070	0.333	0.015	0.029	0.000	0.000	0.000
AfriBERTa_large_supervised	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AfriBERTa_semi_supervised	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 7: **Model Performance Evaluation on Emotion classification Tasks.** *Performance metrics of supervised and semi-supervised learning for (Neutral, Frustration, and Happy) emotion analysis across models. Showing poor performance of happy emotions.*

Model	Anger			Love			Fear		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Distilbert_supervised	0.517	0.861	0.646	0.333	0.222	0.267	0.000	0.000	0.000
Distilbert_semi_supervised	0.445	0.833	0.580	0.357	0.212	0.266	0.000	0.000	0.000
mBERT_supervised	0.524	0.795	0.632	0.408	0.444	0.426	0.000	0.000	0.000
mBERT_semi_supervised	0.487	0.838	0.616	0.438	0.430	0.434	0.000	0.000	0.000
XLM_R_supervised	0.553	0.943	0.697	0.489	0.489	0.489	0.000	0.000	0.000
XML_R_semi_supervised	0.506	0.918	0.652	0.427	0.461	0.443	0.000	0.000	0.000
AfriBERTa_large_supervised	0.484	0.975	0.647	0.250	0.022	0.041	0.000	0.000	0.000
AfriBERTa_semi_supervised	0.417	0.920	0.574	0.182	0.012	0.023	0.000	0.000	0.000

Table 8: **Model Performance Evaluation on Emotion Classification Tasks.** *Performance metrics of supervised and semi-supervised training for (Anger, Love, and Fear) emotion analysis across models. The model performed poorly on Fear emotions.*

Model	Sadness			Empathy			Surprise		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Distilbert_supervised	0.500	0.222	0.308	0.833	0.652	0.732	0.250	0.333	0.286
Distilbert_semi_supervised	0.100	0.083	0.091	0.844	0.580	0.688	0.360	0.337	0.348
mBERT_supervised	0.200	0.222	0.211	0.865	0.660	0.749	0.237	0.500	0.321
mBERT_semi_supervised	0.129	0.167	0.145	0.855	0.621	0.720	0.377	0.516	0.436
XLM_R_supervised	0.250	0.111	0.154	0.791	0.747	0.768	0.000	0.000	0.000
XML_R_semi_supervised	0.154	0.083	0.108	0.767	0.701	0.733	0.250	0.021	0.039
AfriBERTa_large_supervised	0.000	0.000	0.000	0.731	0.719	0.725	0.000	0.000	0.000
AfriBERTa_semi_supervised	0.000	0.000	0.000	0.706	0.659	0.682	0.000	0.000	0.000

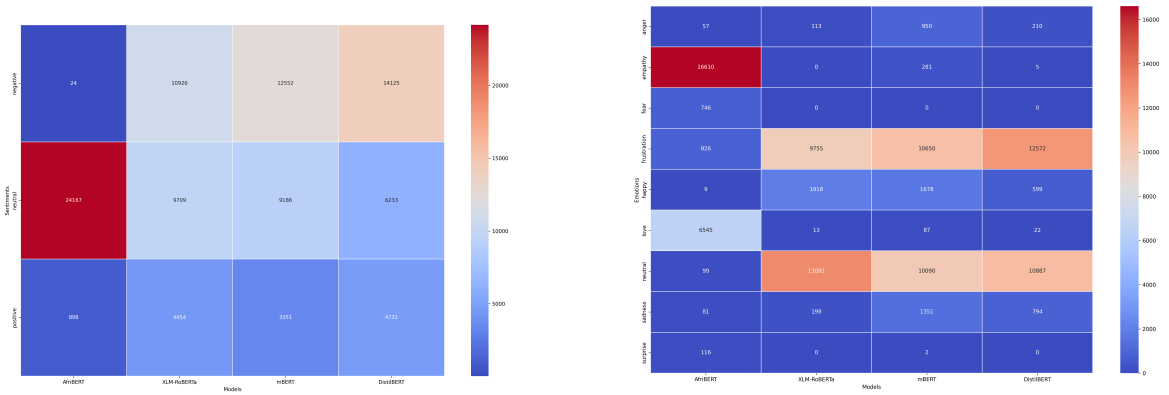
Table 9: **Model Performance Evaluation on Emotion classification Tasks.** *Performance metrics of supervised and semi-supervised training methods for emotion (Sadness, Empathy, and Surprise) analysis across various models. Showing outstanding performance on Empathy emotions.*

presented in Table 6. For the negative sentiment, the supervised XLM-R achieves a high F1-score of 69.9%, unlike the semi-supervised AfriBERTa, which has the worst F1-score of 17.4%. In neutral sentiment classification, the supervised XLM-R again excels with an F1-score of 52.6%. For the positive sentiment, the semi-supervised XLM-R stands out with an exceptional F1-score of 77.1%, and the semi-supervised AfriBERTa shows robust performance with an F1-score of 66.3%.

5.2 Emotion Analysis

Table 5 summarizes the performance of all models on emotion analysis. The models generally show lower performance than sentiment analysis. The supervised DistilBERT achieves the highest F1-score of 31%, followed by mBERT semi-supervised, with an F1-score of 29.7%.

Table 7 shows performance for emotion classification across neutral, frustration, and happy. DistilBERT supervised leads in frustration with an F1-score of 40%. All models perform poorly on happy emotion classification. In Table 8, XLM-R supervised leads for anger and love emotions with F1-scores of 69.7% and 48.9%, respectively, but all models struggle with fear emotion. Table 9 shows low performance for sadness and surprise but outstanding performance for empathy with XLM-R supervised, leading with an F1-score of 76.8%.



(a) Sentiment Prediction Comparison Across Models

(b) Emotion Prediction Comparison Across Models

Figure 4: *Heatmaps comparing sentiment and emotion predictions across different models. AfriBERT model most frequently predicts neutral sentiment and shows the highest sensitivity for empathy emotions.*

5.3 Pretrained Models performance

As shown in Figure 4, XLM-R, particularly in its supervised form, consistently outperforms other models across sentiment and emotion analysis tasks. mBERT also performs reliably well in sentiment analysis and some emotion classifications. DistilBERT, while efficient, has limitations in handling a range of emotions. AfriBERTa shows lower performance across most metrics than other models. Despite being tailored to African languages, AfriBERTa models do not perform as well in sentiment and even worse in emotion analysis.

5.4 Semi-Supervised Performance Analysis

The detailed analysis of SSL models reveals mixed outcomes, with clear performance enhancements in certain models and tasks, particularly in sentiment analysis. For example, mBERT's semi-supervised version slightly improved sentiment analysis with an F1-score of 59.8% compared to 59.6% for supervised version. In emotion analysis, mBERT's semi-supervised version outperformed its supervised counterpart with an F1-score of 29.7% versus 26.5%. The semi-supervised AfriBERTa achieved an F1-score of 36.6% in sentiment analysis, marginally higher than the supervised version's 35.8%, and scored 15.7% compared to 14.2% in emotion task.

6 Limitations

We acknowledge the subjective nature of sentiment and emotion analysis, which can be influenced by label bias, leading to inconsistencies in labeled data. We will publicly share our dataset to address this issue and facilitate further study on label bias and annotator disagreement. Secondly, the cost of obtaining labeled datasets, particularly from native speakers, can be challenging. Transformer models, SOTA for sentiment and emotion analysis, require large data and computational resources, which is still challenging in low-resource setting. Lastly, We recognize the ethical considerations of LLM use.

7 Conclusions and Future Work

We presented RideKE, a code-switched dataset from Twitter, with sentiment and emotion labels partially annotated for Kenyan-accented English mixed with Swahili and Sheng. Our semi-supervised learning shows mixed results, with clear performance enhancements in certain models and tasks, particularly in sentiment analysis, suggesting its potential to generally enhance model performance. We highlight the benefits of semi-supervised learning in improving model performance and reducing data annotation costs.

In the future, we aim to further enhance model performance by expanding the pool of human-labeled datasets, use other semi-supervised approaches, utilizing techniques like few-shot learning, and experimenting with different model architectures and hyperparameters tuning.

Acknowledgments

We thank the volunteer annotators who dedicated their time and expertise to this project, which would not have succeeded without their commitment.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. *arXiv preprint arXiv:2203.08351*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Jason Angel, Segun Taofeek Aroyehun, Antonio Tamayo, and Alexander Gelbukh. 2020. NLP-CIC at SemEval-2020 task 9: Analysing sentiment in code-switching language using a simple deep-learning classifier. *arXiv preprint arXiv:2009.03397*.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023. Sentiment analysis across multiple african languages: A current benchmark. *arXiv preprint arXiv:2310.14120*.

ONAN Aytuğ. 2018. Sentiment analysis on twitter based on ensemble of psychological and linguistic feature sets. *Balkan Journal of Electrical and Computer Engineering*, 6(2):69–77.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204.

JoAnne Banks-Wallace. 2002. Talk that talk: Storytelling and analysis rooted in african american oral tradition. *Qualitative health research*, 12(3):410–426.

Sandra Barasa. 2016. Spoken code-switching in written form? manifestation of code-switching in computer mediated communication. *Journal of Language Contact*, 9(1):49–70.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23.

Kasturi Bhattacharjee, Miguel Ballesteros, Rishita Anubhai, Smaranda Muresan, Jie Ma, Faisal Ladhak, and Yaser Al-Onaizan. 2020. To BERT or not to BERT: Comparing task-specific and task-agnostic semi-supervised approaches for sequence tagging. *arXiv preprint arXiv:2010.14042*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Jan Carter-Black. 2007. Teaching cultural competence: An innovative strategy grounded in the universality of storytelling as depicted in african and african american storytelling traditions. *Journal of Social Work Education*, 43(1):31–50.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. 2006. *Introduction to Semi-Supervised Learning*. MIT press.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Brenda Danet and Susan C Herring. 2007. *The multilingual Internet: Language, culture, and communication online*. Oxford University Press.

Sanjiv Ranjan Das and Mike Y Chen. 2001. Yahoo! for amazon: Sentiment parsing from small talk on the web. *For Amazon: Sentiment Parsing from Small Talk on the Web (August 5, 2001). EFA*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pages 519–528.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ye-Qian Du, Jie Zhang, Xin Fang, Ming-Hui Wu, and Zhou-Wang Yang. 2023. A semi-supervised complementary joint training approach for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Amitabh Vikram Dwivedi. 2014. Linguistic realities in Kenya: A preliminary survey. *Ghana Journal of Linguistics*, 3(2):27–34.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of POS-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–592.

Rahul Gupta, Saurabh Sahu, Carol Espy-Wilson, and Shrikanth Narayanan. 2018. Semi-supervised and transfer learning approaches for low resource sentiment classification. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5109–5113. IEEE.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Hohyun Hwang and Younghoon Lee. 2021. Semi-supervised learning based on auto-generated lexicon using XAI in sentiment analysis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 593–600.

Jahanzeb Jabbar, Iqra Urooj, Wu JunSheng, and Naqash Azeem. 2019. Real-time sentiment analysis on e-commerce application. In *2019 IEEE 16th international conference on networking, sensing and control (ICNSC)*, pages 391–396. IEEE.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*.

Fridah Kanana Erastus and Hilda Kebeya. 2018. Functions of urban and youth language in the new media: The case of Sheng in Kenya. *African youth languages: New media, performing arts and sociolinguistic development*, pages 15–52.

Semi-Supervised Learning. 2006. Semi-supervised learning. *CSZ2006. html*, 5.

Sophia Lee and Zhongqing Wang. 2015. Emotion in code-switching texts: Corpus construction and analysis. In *Proceedings of the Eighth SIGHAN workshop on chinese language processing*, pages 91–99.

M Paul Lewis. 2014. Ethnologue: Languages of the world. https://www.sil.org/about/endangered-languages/languages-of-the-world.

Bing Liu. 2022. *Sentiment Analysis and Opinion Mining*. Springer Nature.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

E Matthew. 2018. Peters, mark neumann, mohit iyyer, matt gardner, christopher clark, kenton lee, luke zettlemoyer. deep contextualized word representations. In *Proc. of NAACL*, volume 5.

Alamin M Mazrui. 1995. Slang and code-switching: The case of Sheng in Kenya. *Afrikanistische Arbeitspapiere: Schriftenreihe des Kölner Instituts für Afrikanistik*, (42):168–179.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179.

Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Clara Momanyi. 2009. The effects of'Sheng'in the teaching of Kiswahili in Kenyan schools. *Journal of Pan African Studies*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, et al. 2023a. Afrisenti: A twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M Mohammad, and Meriem Beloucif. 2023b. Semeval-2023 task 12: sentiment analysis for african languages (afrisenti-semeval). *arXiv preprint arXiv:2304.06845*.

Sebastian Nagel. 2018. Common Crawl - Blog - Index to WARC Files and URLs in Columnar Format — commoncrawl.org. https://commoncrawl.org/blog/index-to-warc-files-and-urls. [Accessed 27-06-2024].

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. 2019. Robustness to adversarial perturbations in learning from incomplete data. *Advances in Neural Information Processing Systems*, 32.

Usman Naseem and Katarzyna Musial. 2019. Dice: Deep intelligent contextual embedding for twitter sentiment analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 953–958. IEEE.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, pages 70–77.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Thomas Niesler and Febe De Wet. 2008. Accent identification in the presence of code-mixing. In *Odyssey*, page 27.

Thomas Niesler et al. 2018. A first south african corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Alvaro Ortigosa, José M Martín, and Rosa M Carro. 2014. Sentiment analysis in facebook and its application to e-learning. *Computers in human behavior*, 31:527–541.

Billian Khalayi Otundo and Martine Grice. 2022. Intonation in advice-giving in kenyan english and kiswahili. *Proceedings of Speech Prosody 2022*, pages 150–154.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Viet H Pham, Thang M Pham, Giang Nguyen, Long Nguyen, and Dien Dinh. 2023. Semi-supervised neural machine translation with consistency regularization for low-resource languages. *arXiv preprint arXiv:2304.00557*.

Mario Piergallini, Rouzbeh Shirvani, Gauri Shankar Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in swahili-english language data. In *Proceedings of the second workshop on computational approaches to code switching*, pages 21–29.

Shana Poplack. 2000. Toward a typology of code-switching. *L. WEI (éd.), The bilingualism reader. London, New York: Routeledge*, pages 221–255.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Fuji Ren and Changqin Quan. 2012. Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. *Information Technology and Management*, 13:321–332.

Takaaki Saeki, Heiga Zen, Zhehuai Chen, Nobuyuki Morioka, Gary Wang, Yu Zhang, Ankur Bapna, Andrew Rosenberg, and Bhuvana Ramabhadran. 2023. Virtuoso: Massive multilingual speech-text joint semi-supervised learning for text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121.

Carol Myers Scotton. 1993. *Social motivations for codeswitching: Evidence from Africa*. Clarendon Press.

Salam Michael Singh and Thoudam Doren Singh. 2022. Low resource machine translation of English–Manipuri: A semi-supervised approach. *Expert Systems with Applications*, 209:118187.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Michelle Terblanche, Kayode Olaleye, and Vukosi Marivate. 2024. Prompting towards alleviating code-switched data scarcity in under-resourced languages with gpt as a pivot. *arXiv preprint arXiv:2404.17216*.

S Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2382–2388. IEEE.

Samuel Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6704–6708. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bao-Khanh Ho Vo and NIGEL Collier. 2013. Twitter emotion analysis in earthquake situations. *Int. J. Comput. Linguistics Appl.*, 4(1):159–173.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-fourth International Joint Conference on Artificial Intelligence*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing systems*, 32.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, et al. 2024. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *arXiv preprint arXiv:2212.09660*.

H Zamani, A Abas, and MKM Amin. 2016. Eye tracking application on emotion analysis for marketing strategy. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(11):87–91.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Appendix

A.1 Language Detection

Language Code	Occurrences	Language
en	29845	English
id	3288	Indonesian
sw	624	Swahili
no	192	Norwegian
da	119	Danish
tr	95	Turkish
nl	81	Dutch
af	73	Afrikaans
de	71	German
ca	55	Catalan
so	46	Somali
sv	34	Swedish
et	26	Estonian
tl	15	Tagalog (Filipino)
hu	14	Hungarian
fr	14	French
es	10	Spanish
hr	9	Croatian
it	8	Italian
cy	8	Welsh
fi	6	Finnish
pl	4	Polish
sl	3	Slovenian
lt	3	Lithuanian
ro	3	Romanian

Table 10: Count of language detection in the RideKE dataset

A.2 Tweets Per Location

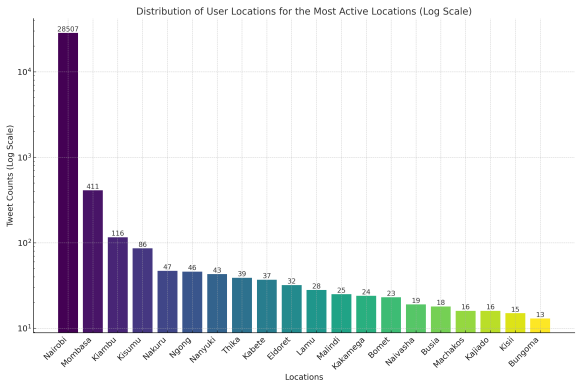


Figure 5: Number of tweets per location on a logarithmic scale. Nairobi appears to be the most active location per dataset.

A.3 Sheng-to-English Sample Sentences

Sheng	English Translation
dere anadai	Driver demands
kuna some people eating	people benefitting
ferry slay queens	Ferry divas
Mmemulikwaa oya	on the spotlight !
Mhesh	honorable sir
wazungu's	white people
sikwembe ya Yesu	strong faith in Jesus
Hiyo pesa ni kadonye	That's little money
fare noma	Expensive fare
kuweka ngata	To fuel

Table 11: Sheng to English Example Sentences

A.4 Annotation Guidelines

Aspect	Details
Title	Annotation Guidelines for RHS Conversation on Twitter
Task	Annotating emotions in tweets related to RHS experiences
Annotation Process	• Emotion Definition: Annotators accurately identify and label the predominant emotion expressed in each tweet based on the emotional tone conveyed by the text. • Keyword Identification: Pay attention to keywords or phrases that suggest the presence of a particular emotion. • Context Matters: Consider the tweet's context, including any relevant hashtags, mentions, or user profiles, for a better understanding of the emotional context. • Tweet Length: Emotions can be expressed differently in short and long tweets.
Emotion Labels Guidelines	1. **Anger:** Label when the tweet expresses frustration, annoyance, resentment, or strong displeasure toward RHS, drivers, or related issues. Look for keywords and tone indicative of anger. Keywords: angry, furious, annoyed, upset. Example: "Terrible experience with Uber driver! He was rude and refused to follow the GPS directions #Angry". 2. **Happy:** Label when the tweet reflects joy, satisfaction, contentment, or delight regarding RHS experiences. Look for expressions of happiness, appreciation, or positive feedback. Keywords: happy, delighted, thrilled, satisfied. Example: "Just had the best ride ever with the friendliest driver! #HappyCustomer #GreatService" 3. **Fear:** Label when the tweet expresses anxiety, worry, concern, or fear about RHS safety, incidents, or perceived risks. Identify cues of fear or apprehension. Keywords: afraid, scared, worried, nervous. Example: "My ride is taking an unfamiliar route, and I'm getting worried. Is this safe? #Fear" 4. **Suprise:** Label when the tweet indicates astonishment, amazement, or unexpected reactions to RHS experiences.Keywords: surprised, shocked, amazed, unexpected. Example: "Wow, my driver gave me a free upgrade to a luxury car! #Surprised 5. **Love:** Label when the tweet reflects affection, appreciation, or strong positive emotions toward RHS, drivers, or related aspects. Look for expressions of love or admiration. Keywords: love, adore, appreciate, grateful. Example: "Wow, my driver gave me a free upgrade to a luxury car! #Surprised #Love" 6. **Frustration:** Label when the tweet expresses dissatisfaction, irritation, or being fed up with RHS issues. Identify cues of frustration and annoyance. Keyword: frustrated, annoyed, fed up, irritated. Example: "Been waiting for my ride for ages. This is so frustrating!#Frustrated #LateAgain" 7. **Neutral:** Label when the tweet does not exhibit any strong emotional sentiment or when the emotion is unclear or ambiguous. Use this label sparingly and only when other emotions are not evident. Example: "Just booked my ride for tomorrow morning. #RideHail #PlanningAhead
Quality Control	Monitor inter-annotator agreement to ensure consistency among annotators. Resolve disagreements through discussion and clarification.
Privacy and Ethical Considerations	Respect user privacy and report any offensive content appropriately.

Table 12: Annotation guidelines for ride-hailing service conversation emotions on Twitter

A.5 Sample dataset structure

Keyword	Date	Tweets	reply count	retweet count	like count	verified	user followers	user following	user tweets	user location	country
#UBER-Kenya	2023-04-10	Did Nairobi ask you to double Nairobi fare price ? That's how Uber Kenya and bolt steal from us here.	1	0	0	0	2104	981	23173	Mombasa	Kenya
#UBER-Kenya	2023-03-30	Uber Kenya made an order that was cancelled by a restaurant but I've already paid. How do I follow up on my refund?	1	0	0	0	946	975	4642	Nairobi	Kenya
#UBER-Kenya	2023-03-30	Uber Kenya made an order that was cancelled by a restaurant but I've already paid. How do I follow up on my refund?	1	0	0	0	946	975	4642	Nairobi	Kenya
#UBER-Kenya	2023-04-02	Uber is losing the Kenyan market to Nairobi apps, customers are tired of being asked by drivers where in Nairobi they are going. Nairobi apps show Nairobi drivers where the customer is, where is going and price hence drivers will decide to accept or decline the request.	2	0	0	0	46	297	817	Nairobi	Kenya
#UBER-Kenya	2023-03-27	Uber Kenya how can your driver click not paid when he was paid? And Nairobi is proof of payment?	2	0	0	0	5744	1338	208846	Nairobi	Kenya
#BOLT-Kenya	2023-04-06	Hello, thanks for writing in. Kindly do reach out to us via kenyabolt.eu and a member of our team will respond and assist accordingly.	0	0	0	1	15093	447	16966	Nairobi	Kenya
#BOLT-Kenya	2023-01-16	Let's have an honest conversation here...this morning you lowered the base category to 8ksh per kilometer. We all know that fuel is still very high. What method did you use to reach this point, Did you involve drivers about the	1	0	3	0	14	87	82	Nairobi	Kenya
#BOLT-Kenya	2022-11-19	If you don't communicate. Let us as drivers do what we feel like doing. Because bolt Kenya is manner less.	0	0	0	0	11	69	66	Nairobi	Kenya
#BOLT-Kenya	2019-10-27	How come Bolt Kenya does not have an active customer service line for queries?	0	0	0	0	23	180	35	Nairobi	Kenya
#BOLT-Kenya	2019-05-20	Thanks to Boltkenya been arriving at my studio sessions and interviews on time and with comfort. You too can enjoy this service by simply downloading boltkenya and using my code FEMIONEBOLT to get kshs250 off	0	0	3	0	40629	528	24042	Nairobi	Kenya
#LITTLECAB	2022-07-31	Now #Littlecab will not allow me to cancel a ride I did not take until I pay. Exhausting!	0	0	0	0	636	2222	4085	Nairobi	Kenya
#LITTLECAB	2022-07-31	And they let me have their driver. The security officer at #Carnivorekenya says that they do not verify the drivers. What is the whole point of telling us to use #littlecab if you have no relationship with them. Just destroyed my whole experience attending a beautiful musical.	1	0	0	0	636	2222	4085	Nairobi	Kenya
#LITTLECAB	2022-05-10	Use #Littlecab. These other Apps are foreign and exploitive.	0	0	1	0	480	987	4740	Mombasa	Kenya
#LITTLECAB	2020-12-23	Why do we always encounter cabs from #LittleCab that arrive with different number plates from what is registered in your system? While I don't board them in principle for security concerns, it may one day be costly for a desperate client	2	0	0	0	4712	3044	20683	Nairobi	Kenya

Table 13: Original sample of the tweets data structure