Analysing Public Transport User Sentiment

Rozina L. **Myoya**

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Faculty of Engineering, Built Environment & IT,
Department of Computer Science, University of Pretoria, Pretoria.

Analysing Public Transport User Sentiment

Supervised by
1st Supervisor - Prof. Vukosi **Marivate**
2nd Supervisor - Dr. Idris **Abdulmumin**

April 2, 2024

Declaration

I, Rozina Myoya, hereby declare the content of this dissertation to be my own work unless otherwise explicitly referenced. This dissertation is submitted in partial satisfaction of the requirements for a Masters degree in Big Data Science at the University of Pretoria, Pretoria. This work has not been submitted to any other university, nor for any other degree.

Signed: _____

Date: _____

Abstract

In many Sub-Saharan countries, the advancement of public transport is frequently overshadowed by more prioritised sectors, highlighting the need for innovative approaches to enhance both the Quality of Service (QoS) and the overall user experience. This research aimed at mining the opinions of commuters to shed light on the prevailing sentiments regarding public transport systems. Concentrating on the experiential journey of users, the study adopted a qualitative research design, utilising real-time data gathered from Twitter to analyse sentiments across three major public transport modes: rail, mini-bus taxis, and buses. By employing Multilingual Opinion mining techniques, the research addressed the challenges posed by linguistic diversity and potential code-switching in the dataset, showcasing the practical application of Natural Language Processing (NLP) in extracting insights from under-resourced language data. The primary contribution of this study lies in its methodological approach, offering a framework for conducting sentiment analysis on multilingual and low-resource languages within the context of public transport. The findings hold potential implications beyond the academic realm, providing transport authorities and policymakers with a methodological basis to harness technology in gaining deeper insights into public sentiment. By prioritising the analysis of user experiences and sentiments, this research provides a pathway for the development of more responsive, user-centered public transport systems in Sub-Saharan countries, thereby contributing to the broader objective of improving urban mobility and sustainability.

Acknowledgements

I wish to express my sincere gratitude to the following organizations and individuals, whose contributions were invaluable to the successful completion of this dissertation:

a) Professor V. Marivate, my supervisor, and Dr. Idris Abdulmumin, my co-supervisor, for their continuous guidance and support throughout the research process.

b) The Data Science for Social Impact (DSFSI) research group, for providing access to their annotated datasets and ongoing support.

c) The Masakhane group, for their invaluable input and assistance in data sourcing.

d) The numerous anonymous Twitter users, whose data and opinions were essential in enabling this research.

e) My family and friends, for their unwavering encouragement and support throughout my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Developing a comprehensive user experience map of the public transportation system, which is a crucial component of the everyday lives of millions of users, offers the chance to better understand user behavior and spot opportunities that can result in a rise in public transport usage. The use of social media platforms has evolved into a viable source of information on the user experience of public transportation because commuters use these platforms to share their comments or complaints on a wide range of topics [Batorsky, 2022; Cndro, 2021; Haghighi et al., 2018]. Transit authorities have recently started using social media in the public transportation sector to enhance user interaction. For instance, Cndro [2021]noted that transit authorities use social media as a way of engaging with the public, providing timely updates, alerting the public about emergencies, and promoting the usage of public transportation.

Social media platforms have become vital tools for mining public opinions in various sectors, including public transportation. Researchers have leveraged these platforms to understand the sentiments and preferences of commuters, offering insights that can significantly enhance service delivery. For instance, Murçós [2021] utilised social media data to analyse urban mobility patterns, identifying common grievances and satisfaction levels among public transport users . This approach enabled the identification of critical areas for improvement in urban transport systems. Similarly, Qi et al. [2020] developed a comprehensive framework to analyse tweets related to public transport services. Their methodology not only highlighted the predominant negative sentiments among commuters, often related to service delays and discomfort but also showcased the potential of positive feedback in identifying well-received service features [Qi et al., 2020]. These studies underscore the power of social media analytics in capturing the diverse opinions of commuters, offering invaluable data that can inform policy and operational adjustments in the public transport sector.

This research explored the application of Natural Language Processing (NLP) in the multilingual domain and aimed to open the door to another avenue of using technology to gain insight into the commuter experience. The aim was to find spaces within the public transport system, specifically targeting bus, train, and mini-bus taxi modes, where the user experience could be

enhanced despite the constraints of implementing large-scale interventions. To determine these pockets of opportunity, information mining and sentiment analysis was carried out on commuter experiences extracted from Twitter. Given the expected multilingual nature of the data, the study explored methods of carrying out multilingual opinion mining and transfer learning [Ekinci et al., 2018; Kuratov and Arkhipov, 2019]. The study employed existing datasets in English, Swahili, isiZulu, and SeTswana for sentiment analysis, as detailed in the works of Mabokela and Schlippe [2022]; Mambina et al. [2022]; Muhammad et al. [2023b,c].

This research exemplifies the practical application of multilingual sentiment analysis for in-depth market insights in public transport. Our primary objective was to determine the core issues faced by commuters, thereby gaining a deeper understanding of user experiences in public transportation. We also sought to ascertain whether these user-identified challenges align with the problem areas recognised by public transport providers. The successful implementation of this research could lead to further application of NLP to determine the end-user experience in domains such as financial institutions, healthcare facilities, etc. that serve communities that communicate in languages other than English. The inclusion of the underrepresented creates an opportunity for new sources of invaluable information regarding the African market and user behavior, while also enriching existing datasets of under-resourced languages to include the transport domain.

1.2 Problem Statement

The user experience of public transport in sub-Saharan countries often receives inadequate attention, leading to a significant discrepancy between the perceptions of commuters and service providers. This gap in understanding results in services that fail to cater to the actual needs of users [Batorsky, 2022; Vicente and Reis, 2016]. Predominantly, service providers adopt a "one-size fits all" approach, which, while convenient, falls short in appealing to a broader user base. Consequently, these services are frequently the option of necessity rather than choice, primarily utilised by captive users who have no alternatives [Camacho et al., 2016].

Investment in public transport systems in many sub-Saharan countries is either on the decline or entirely absent, posing a formidable challenge to conducting comprehensive market research and executing substantial system improvements [Camacho et al., 2016]. Our research seeks to bridge this gap. By harnessing the power of opinion mining and sentiment analysis, we can uncover valuable insights into the commuter experience. These insights offer opportunities for targeted enhancements that are both cost-effective and impactful.

The goal is not just to increase the overall ridership of public transport but to significantly enhance the quality of service for existing users. By strategically analysing commuter sentiments and feedback, we can pinpoint specific areas requiring improvement. Implementing these changes could lead to a more satisfying and user-centric public transport experience, potentially attracting new users and enriching the journey for regular commuters. This approach represents a shift towards data-driven decision-making in public transport, where user opinions play a pivotal role in shaping service improvements.

Our study aimed to address the following research questions:

1. **Understanding User Sentiment in Public Transport:**
    - How do commuters perceive and experience public transport, as reflected in their real-time sentiments on Twitter?
    - What specific aspects of public transport are most frequently discussed by users in these regions, and what sentiments (positive, negative, neutral) do they express?
2. **Evaluating Experiences Across Transport Modes:**
    - How do user experiences and sentiments differ across various modes of public transport, such as rail, mini-bus taxis, and buses?
    - Are there noticeable trends or patterns in user satisfaction or dissatisfaction specific to each transport mode?
3. **Application of Multilingual Opinion Mining:**
    - How can multilingual opinion mining techniques effectively handle the challenges of language diversity and code-switching in social media datasets from Sub-Saharan countries?
    - What insights can be gained about public transport from code-mixed and multilingual data that traditional monolingual analysis might miss?
4. **Comparative Analysis of Service Ratings and User Sentiments:**
    - How do user sentiments on public transport, as expressed on social media, compare with official service ratings provided by transport providers?
    - What discrepancies, if any, exist between the perceived quality of service from providers and the actual experiences of users?
5. **Practical Use of NLP in Enhancing Public Transport Systems:**
    - How can Natural Language Processing (NLP) be practically applied to extract valuable insights from under-resourced language data in the context of public transport?

1.3 Scope of the Study

This study was focused on analysing data collected from Twitter through specific keyword searches related to public transport, specifically targeting bus, train, and mini-bus taxi modes. The time frame for data collection spanned from 1^{st} of January, 2018, to 1^{st} of March, 2023. Within this scope, the analysis was limited to four languages: English, isiZulu, SeTswana, and Swahili. The primary focus of the insights and data extraction was on aspects relevant to public transport.

1.4 Methodology

This study was focused on implementing comprehensive sentiment analysis on Tweets related to public transport sourced from Kenya, South Africa, and Tanzania. Given the multilingual nature of the dataset, which included a blend of English, isiZulu, SeTswana, and Swahili, special attention was devoted to the analysis of code-mixed data. Additionally, machine learning models pre-trained in these languages were employed to ensure accurate sentiment analysis across the diverse linguistic spectrum. The methodology for data extraction, processing, and analysis was systematically structured into five pivotal steps: (i) Data sourcing; (ii) Data Pre-Processing and Exploratory Data Analysis; (iii) Model Development and Evaluation; and lastly, (iv) Comparison and Validation of Sentiment Predictions against Public Transport Provider Ratings.

(i) **Data Sourcing:** The data collection process involved scraping Twitter for specific keywords related to public transport targeting each of the three countries. The timeframe for this data collection spanned from 1^{st} January 2018 to 1^{st} March 2023. Adhering to ethical standards and prioritising user privacy, all personally identifiable information, including usernames and location tags, was meticulously removed from the dataset to ensure the complete anonymity of the Twitter users whose data was analysed.

(ii) **Data Pre-Processing and Exploratory Data Analysis:** This phase entailed a thorough investigation to identify any trends in the data based on timelines, keywords, or other relevant factors. Given the multilingual composition of the dataset, language identification was executed, and the distributions of different languages were determined. This stage also involved feature extraction to derive deeper insights into the dataset. This phase also included the detailed descriptions of the annotated datasets used in this study.

(iii) **Model Development and Evaluation:** This critical step encompassed the evaluation of the chosen pre-trained language models (PLMs). The models were assessed based on their performance using the annotated datasets. The study also examined the efficacy of a Siamese Network in addressing the prevalent code-mixed data, involving both the development and evaluation of this model.

(iv) **Comparison and Validation of Sentiment Predictions:** Finally, sentiment predictions were made using the best-performing PLMs, and these sentiments were juxtaposed with public transport provider ratings to evaluate alignment. The results were then validated to ascertain the models' reliability in predicting commuter sentiment.

1.5 Organisation of the Report

The layout of the dissertation is as follows:

- Chapter 1 serves as the introduction to the dissertation summarising the motivation behind the study, the scope of the study, and the methodology followed.
- Chapter 2 outlines the literature review on the research topic and other related topics that contributed to the research.

- Chapter 3 outlines the methodology followed in detail and provides the results and insights obtained from the analyses carried out.
- Chapter 4 outlines the model development and evaluation carried out and provides the results and the conclusions drawn from the results obtained.
- Chapter 5 concludes the dissertation with findings of the research and recommendations.

Chapter 2

Literature Survey

2.1 Introduction

Traditionally, insights into user experiences within public transport systems have been obtained through extensive survey methods, including public participatory forums, questionnaire surveys, and observational studies utilising recordings, observers, or GPS tracking [Casas and Delmelle, 2017; Keseru et al., 2020; Rieser-Schüssler and Axhausen, 2013]. While these conventional methods are effective, they demand significant time and resources to gather comprehensive data. The rise of social media as a platform for sharing opinions and grievances offers new, dynamic avenues for data collection. Platforms like Twitter have become increasingly recognised as rich sources of real-time public sentiment on various topics, including public transport [Casas and Delmelle, 2017; Cndro, 2021; Haghighi et al., 2018].

The shift towards mining opinions from social media brought about the challenge of creating and accessing resources capable of effectively analysing social media data, which often includes colloquial, noisy, and code-mixed content [Ledwaba and Marivate, 2022]. To effectively execute Natural Language Processing (NLP) tasks like sentiment analysis, there is a critical need for language resources, particularly annotated datasets. These datasets are indispensable for training and evaluating models to ensure that they yield reliable and representative outcomes. High-quality (or 'gold standard') datasets are essential to ensure trustworthy results in order to avoid the computing phenomenon of 'garbage in, garbage out' [Kim et al., 2016].

After the creation of high quality datasets comes the need for the selection of the appropriate language model to be used. A language model is an NLP tool developed to process, analyse, and generate natural language, trained using methods from rule-based to deep learning. Its applications include text completion, text-to-speech conversion, language translation, and powering chatbots and virtual assistants [OECD, 2023]. These models can either be developed from scratch or existing models can be fine-tuned to suit the required task. The former approach often demands significant investment in terms of processing power and time, whereas the latter offers easy access and immediate implementation, especially beneficial when resources are limited [Alabi et al., 2022]. Pre-trained models carry the advantage of having undergone extensive initial training; however, fine-tuning these models to specific tasks and datasets presents its own

set of challenges [Alabi et al., 2022].

The application of these models to specific use cases, such as sentiment analysis in this study, brings into focus the relevance and applicability of the results obtained. In real-world scenarios, it is vital to assess the reliability of these models, especially considering the unique characteristics of the focus countries in this study. This assessment ensures that the outcomes derived are both reliable and pertinent to the practical applications intended.

2.2 The Use of Social Media in Opinion Mining

In recent years, the use of social media for opinion mining has become increasingly prevalent, offering a rich reservoir of real-time, user-generated data. This transition from traditional survey methods to digital platforms has revolutionised the way insights into public opinion are gathered [Casas and Delmelle, 2017]. Social media platforms, with their diverse and active user bases, present a unique opportunity to capture a wide range of opinions, trends, and sentiments across various demographic and geographic segments. As a dynamic and interactive medium, social media allows for the collection of spontaneous and candid feedback on a multitude of topics [Casas and Delmelle, 2017; Kuflik et al., 2017].

The landscape of social media usage has transformed dramatically over recent years within the continent. WhatsApp, for instance, dominates the market with nearly 90 percent of the continent's social media users, closely followed by giants like Facebook and Twitter [Statista, 2023]. This trend prompts a pertinent inquiry into the viability of harnessing data from these platforms for user experience analysis in the public transport sector. While these platforms offer an abundant source of data, leveraging information from private channels like WhatsApp groups introduces complex issues surrounding data privacy and consent. Factors such as participant incentives, securing informed consent, and ensuring data anonymisation play a crucial role in determining the feasibility and ethicality of utilising chat data for research purposes. Consequently, platforms like Facebook and Twitter, which offer a balance between data availability and adherence to consent and privacy norms, emerge as more practical options for data sourcing in research contexts [Kohne et al., 2022]. These platforms provide a rich repository of user-generated content that is invaluable for understanding public sentiment, particularly in sectors like transportation, where user experience is a key metric for service evaluation and improvement.

The integration of social media in the public transport sector has gained momentum, with transit agencies increasingly utilising these platforms to enhance their interaction and engagement with users. According to Cndro [2021], transit agencies are leveraging social media for a variety of purposes including public education and awareness, active engagement with the community, disseminating quick updates, providing critical information during crises, and promoting the use of public transport. A notable instance of this trend is the #adoptastation campaign spearheaded by PRASA (Metrorail, South Africa) similar to the community engagement initiative documented by Alexander [2012], aimed at promoting the rehabilitation and care of train stations, thereby enhancing public engagement in transport infrastructure maintenance.

This shift towards social media engagement by transit agencies can be viewed as a response to the gradual decline in public transport services within the study's focus countries. Over the

years, there has been a marked deterioration in the quality of public transport, leading to a scenario where the majority of users are captive, relying on public transport due to a lack of alternative options [Clark and Crous, 2002; Vicente and Reis, 2016]. Budgetary constraints and the prioritisation of other critical sectors such as healthcare and education have led to the neglect of public transport systems. Consequently, there is an urgent need for innovative solutions that extend beyond merely improving the Quality of Service (QoS). These solutions should aim to not only enhance the user experience but also actively involve users in the system's continuous improvement.

In addressing these challenges, Clark and Crous [2002] advocates for the adoption of qualitative research methodologies and a multidisciplinary approach. This strategy focuses on understanding and designing solutions centered around the users' experiential journey through the transport system, taking into account their unique perspectives and needs. Such an approach, as suggested by [Mabokela and Schlippe, 2022; Tauchmann, 2021], requires a holistic understanding of the user experience, encompassing both the practical aspects of transportation and the emotional and experiential dimensions. By doing so, public transport agencies can develop more user-centric services, fostering a positive relationship with the commuting public and potentially increasing the overall appeal and usage of public transportation systems.

The results of text mining have real business value in identifying business opportunities within a market. Take the example of Casas and Delmelle [2017] that used opinion mining to identify the points of user dissatisfaction and used that data to inform their next system improvement interventions in a bid to increase BRT ridership. Furthermore, customer segmentation can be conducted to determine the different types of customers that use the service and thus identify population groups that have been excluded from the service or identify commuter groups where there is an opportunity for growth in market share within the service [Ekinci et al., 2018; Krizek and El-Geneidy, 2007]. With the rise in social media applications in our everyday lives, the use of NLP in sentiment analysis is increasingly becoming a sort after application by businesses and the public sector to gauge the end-user experience [Batorsky, 2022; Cndro, 2021; Ekinci et al., 2018; Haghighi et al., 2018; Krizek and El-Geneidy, 2007].

The successful execution of opinion mining, particularly in the context of multilingual data, hinges on the availability of appropriate tools for both language identification and subsequent analysis [Qi et al., 2019]. This is a critical step, especially when dealing with low-resource languages, which traditionally suffer from a lack of effective language identification tools. In recent years, however, there has been a notable advancement in this area with the development of tools such as *AfroLID* [Adebara et al., 2022], *Franc*, and *CLD3* [Salcianu et al., 2018], each with its unique operational framework:

1. **AfroLID** [Adebara et al., 2022]: Distinguished as a pioneering tool for identifying a vast array of African languages, AfroLID stands out with its coverage across 50 African countries. Its main feature, the 'classify' function, efficiently processes and tokenizes input text before feeding it into the model. The function then employs a softmax function to compute language prediction probabilities and returns a detailed dictionary containing language labels, scores, and other relevant information such as the language's name and script. However, a notable limitation of AfroLID is its inability to identify high-resource languages, such

as English. This shortcoming can hinder its effectiveness in capturing the full linguistic diversity within our dataset. Despite this, AfroLID's proficiency in recognising African languages warranted its consideration for our study.

2. **Franc:** Franc is a language detection tool that primarily functions through the analysis of trigrams (sequences of three characters) found in text samples. The script processes multiple languages, each represented by a unique trigram model, stored in a numeric data structure. When given a text input, the script initially determines the top script (writing system) and then evaluates the input against the trigram models of various languages under that script. It calculates a 'distance' for each language, representing how closely the input matches the language's trigram profile. The script considers user-defined parameters like minimum sample length, and whitelist or blacklist languages for more targeted detection. Languages are ranked based on their calculated distances from the input text, with the closest matches being identified as the most probable languages. This system enables the 'franc.py' script to effectively pinpoint the language of a given text sample within its coverage, with special attention to trigram frequency and distribution.

3. **CLD3** [Salcianu et al., 2018]: Google's CLD3 language detection model employs a neural network approach for identifying languages. It begins by extracting character n-grams from the input text and calculating the frequency of each n-gram. These n-grams are then transformed into unique identifiers linked to dense embedding vectors, which are learned during training. The model averages these embeddings based on the n-gram frequencies, and the averaged embeddings form an embedding layer. This layer is processed through a hidden layer with a rectified linear activation function and a softmax layer that outputs probabilities for language prediction. The model's efficacy lies in its ability to process text through this advanced neural network structure, efficiently predicting the language of any given input text.

With the development of the trend of opinion mining through social media, the evolving challenge has now shifted from mere data extraction to ensuring the data mining process encompasses a comprehensive representation of the entire user market. This development highlights the need for methodologies that are adept not only at extracting data but also at accurately interpreting the varied sentiments and opinions expressed by diverse user groups. This need is particularly acute when considering non-English speakers, whose voices often remain unheard because existing opinion mining tools may not adequately cater to low-resource languages.

2.3 Existing Resources for Low Resource Languages

In the context of our study's focus countries – Kenya, South Africa, and Tanzania – language distribution reveals that English is not the predominant mode of communication, as illustrated in Figures 2.1 to 2.3. In South Africa, isiZulu emerges as the most commonly spoken language, both within and outside households, with English being the second most prevalent language for external communication. Similarly, in East Africa, specifically in Kenya and Tanzania, Swahili is the predominant language of communication. This linguistic landscape presents a substantial

reservoir of data in under-resourced languages, yet untapped in the field of Natural Language Processing (NLP).

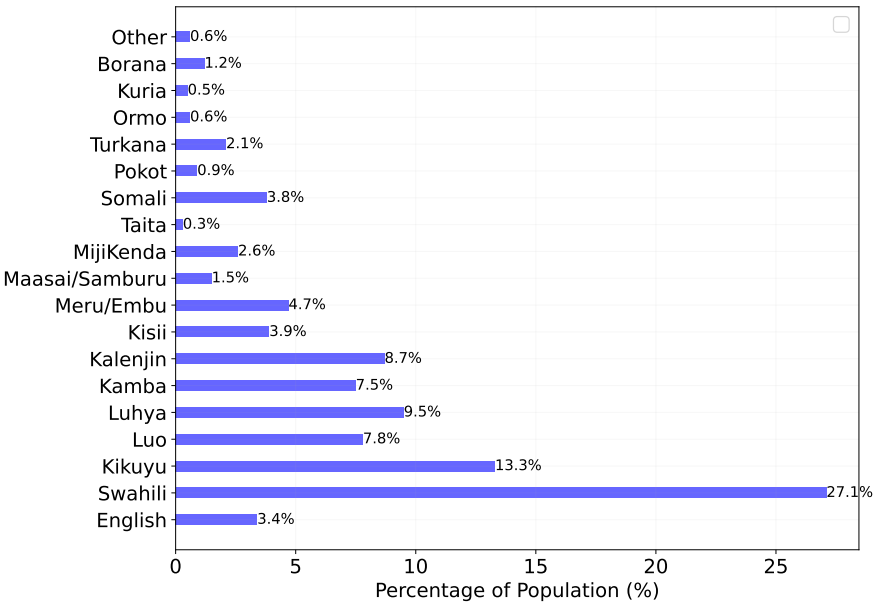

FIGURE 2.1: Percentage distribution of languages spoken in Kenya Statista [2023]

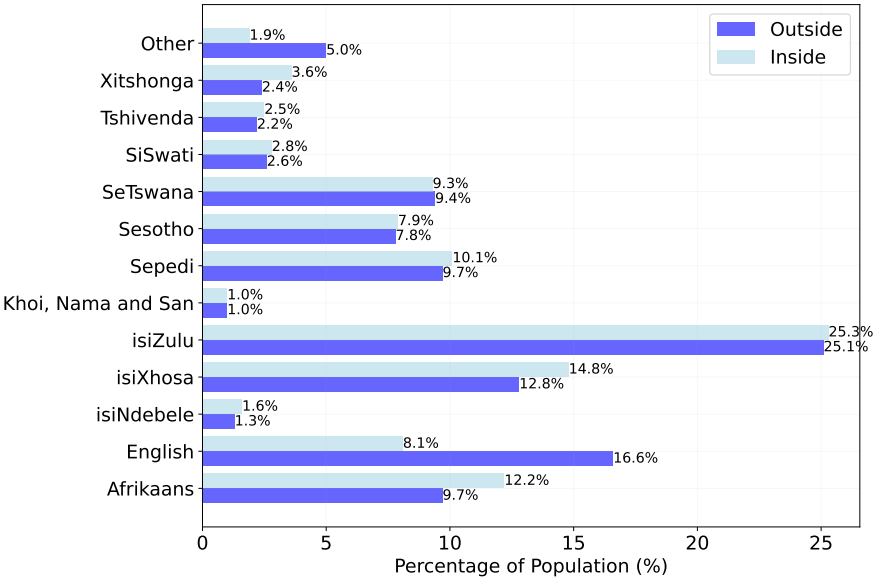

FIGURE 2.2: Percentage distribution of languages spoken in South Africa Statista [2023]

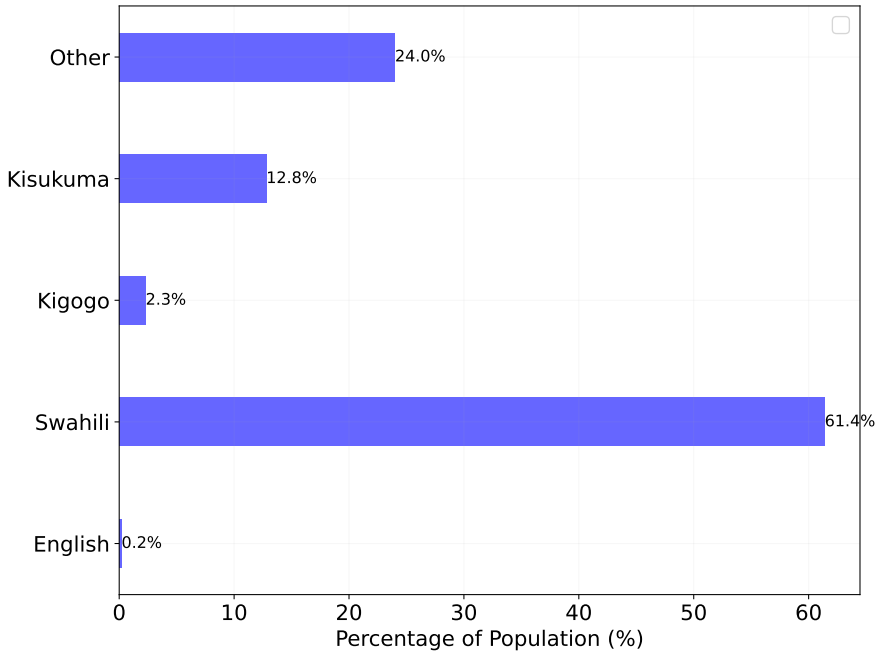

FIGURE 2.3: Percentage distribution of languages spoken in Tanzania Statista [2023]

As depicted in Figure 2.4 and Table 2.1, class five languages, including English, Spanish, German, Japanese, and French, boast the highest number of labeled datasets for NLP applications. However, class 2 languages, such as isiZulu, are still underrepresented in terms of labeled dataset availability. This disparity highlights a critical gap in the NLP domain – the under-representation of linguistic data from low-resourced languages. The language distribution shown in Figures 2.1 to 2.3 underscores this gap and points to the untapped potential in harnessing this rich linguistic data. By expanding the focus to these languages, NLP can offer more inclusive and accurate insights, reflecting a broader spectrum of user sentiments and experiences in the public transport sector and beyond.

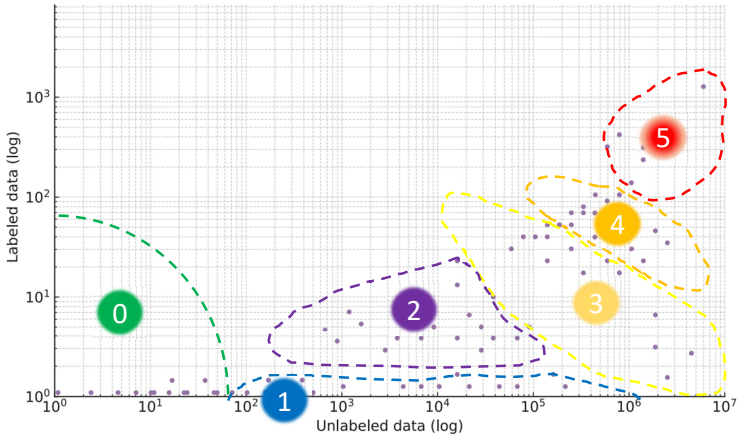

FIGURE 2.4: The comparison of languages used in NLP applications according to labeled vs unlabeled datasets adapted from Batorsky [2022].

TABLE 2.1: Number of languages, speakers, and the percentage of total languages for each language class adapted from Batorsky [2022].

Class	5 Example Languages	Langs	Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca Wallisian, Bora	2191	1.0B	88.17%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	1.0B	8.93%
2	Zulu, Konkani, Lao, Maltese, Irish	19	300M	0.76%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.1B	1.13%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	1.6B	0.72%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Addressing the shortage of multilingual corpora necessitates sourcing and meticulously annotating datasets, a process that is extremely time-consuming and resource-intensive. Various strategies, such as employing expert annotators, crowd-sourcing, or a mix of both, are outlined in Tauchmann [2021] for the annotation of unlabeled data. The study by Choudhary et al. [2018] suggested auto-labelling using prevalent data indicators like emojis for social media-sourced data. However, the challenge lies in developing methods for corpus creation that are not overly time-consuming while ensuring comparable evaluation results across languages [Helmreich et al., 2004]. This issue makes sentiment analysis particularly challenging for under-resourced languages due to the lack of databases for sourcing seed data.

Significant strides in multilingual corpora creation have been made through studies like Muhammad et al. [2023a] and Mabokela and Schlippe [2022], resulting in the creation of the *AfriSenti* and *SAfriSenti* corpora respectively. These include thousands of annotated tweets in numerous African languages and English, with a considerable portion of code-switched texts. Given the prevalence of code-switching in social media within the focus countries, it is crucial to consider this linguistic phenomenon in sentiment analysis. Studies by Choudhary et al. [2018] and Gupta et al. [2016] delve into the challenges of working with code-mixed data and propose novel approaches like using a Siamese Network for sentiment analysis. The increase in social media usage across the African continent necessitates the development of sophisticated methodologies for handling code-mixed data, particularly data emanating from social media platforms.

2.4 The Emergence of Pre-Trained Language Models

The emergence of pre-trained language models (PLMs) has marked a paradigm shift in the field of natural language processing (NLP), introducing an era where complex linguistic tasks are managed with remarkable precision and efficiency [Wang et al., 2022]. These models, which are trained on extensive text corpora, excel in a diverse range of applications, including text generation and sentiment analysis. However, a notable limitation arises when these models are applied to languages excluded from their pre-training datasets, an issue that is particularly acute with African languages [Alabi et al., 2022; Dossou et al., 2022; Ogueji et al., 2021]. These languages are often significantly underrepresented in global language datasets, creating a gap in the effectiveness of PLMs for these languages. This gap has necessitated the development

of innovative solutions like multilingual adaptive fine-tuning, which extends the capabilities of PLMs to low-resource languages by fine-tuning them on these specific linguistic datasets [Alabi et al., 2022]. Despite its effectiveness, this approach requires substantial resources, thereby emphasising the need for models pre-trained specifically on target languages, ensuring a more inclusive and comprehensive approach to language processing [Alabi et al., 2022].

In line with this, the adaptation and refinement of BERT and XLM-R based models have given rise to specialised models such as AfriBERTa [Ogueji et al., 2021], PuoBERTa [Marivate et al.], AfroXLMR [Alabi et al., 2022], and AfroLM [Dossou et al., 2022], which have been pre-trained inclusively to encompass African languages. A succinct description of each model is provided below:

1. **AfriBERTa** [Ogueji et al., 2021]: Specializing in low-resource languages, AfriBERTa is a multilingual language model that encompasses 11 African languages. It has been rigorously evaluated for two pivotal Natural Language Processing (NLP) tasks: text classification and named entity recognition. Notably, in certain scenarios, AfriBERTa has outperformed well-known models like XLM-R and mBERT, underlining its effectiveness in handling diverse African languages.

2. **AfroXLMR** [Alabi et al., 2022]: Developed through an innovative multilingual adaptive fine-tuning process, AfroXLMR includes 3 major high-resource African languages and 17 extensively-resourced African languages. Its evaluation covered a broad range of NLP tasks, including sentiment classification, news topic classification, and named entity recognition, thereby demonstrating its wide-ranging applicability and versatility in handling complex language tasks.

3. **AfroLM** [Dossou et al., 2022]: Built on a groundbreaking self-active learning framework, AfroLM is a multilingual language model trained from the ground up on 23 African languages. Demonstrating superior performance over other multilingual models such as AfriBERTa and mBERT, AfroLM stands out as a robust tool for a variety of linguistic applications, thanks to its comprehensive training approach.

4. **PuoBERTa** [Marivate et al.]: Tailored specifically for Setswana, PuoBERTa is a masked language model that has been trained on a range of monolingual sources, including the NCHLT Setswana corpus, the South African Constitution, Leipzig Setswana BW, ZA corpora, Vuk'zenzele Setswana Corpora, and South African Cabinet Speeches. Its performance in NLP downstream tasks such as part-of-speech tagging, named entity recognition, and news categorization has been commendable. In our study, we particularly explored PuoBERTa's capabilities in sentiment classification, providing a valuable benchmark for its application in this specific area.

The development of these models significantly enhanced the accessibility and applicability of NLP tools, particularly for African languages, enabling critical NLP tasks such as Named Entity Recognition and opinion mining without the need for extensive resource investment in training and development. However, as pointed out in the article by Nicholas and Bhatia [2023], this progress also surfaces socioeconomic challenges stemming from the disproportionate representation of languages within the NLP space. In the process of developing these models, compromises

are often made between languages, leading to a potential trade-off in performance. For instance, optimising a model for Hindi could inadvertently impair its performance in English. Moreover, technology companies may be inclined to prioritise languages spoken by wealthier or more politically influential demographics, potentially exacerbating the existing resource gap for languages that are similar to other low-resourced languages [Nicholas and Bhatia, 2023]. This dynamic poses a significant risk to many African languages, including Swahili, Amharic, and Kabyle, which may find themselves further marginalised in the NLP landscape [Joshi et al., 2020]. This scenario underscores the critical need for equitable and balanced development in language model technology, ensuring that advancements in NLP are beneficial and accessible to all linguistic communities.

This research contributes to this endeavor by applying NLP to sentiment analysis within the public transport sector. It aims to extract meaningful insights from user experiences, potentially uncovering areas for service improvement and increasing ridership. The growing prevalence of social media in our daily lives makes it a rich source of data for sentiment analysis. This approach can provide valuable business intelligence and inform public sector strategies, affirming the timeliness and importance of this study. By focusing on a sector that is vital to the daily lives of millions and leveraging the power of NLP to process multilingual and code-mixed data, this research stands at the forefront of using advanced technology to address real-world challenges, bridging the gap between technology, language, and public service.

2.5 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE, or Synthetic Minority Over-sampling Technique, is an innovative approach designed to address the issue of class imbalance in machine learning datasets. Class imbalance occurs when the number of instances of one class significantly outnumbers those of another, often leading to biased models that favor the majority class. SMOTE effectively combats this by creating synthetic samples of the minority class, thereby balancing the dataset and improving model performance. SMOTE works as follows:

1. **Identifies the Minority Class:** SMOTE initially identifies the minority class that needs over-sampling to balance the class distribution.
2. **Chooses Samples:** It randomly selects a sample from the minority class.
3. **Finds Nearest Neighbors:** For each minority class sample, SMOTE finds its k-nearest neighbors. The number of neighbours (k) is a parameter that can be set by the user.
4. **Synthetic Sample Generation:** For each selected sample, SMOTE generates synthetic instances. This is done by choosing one of the k-nearest neighbours and then creating a new sample that is a linear interpolation between the selected sample and its chosen neighbour.
5. **Repeating the Process:** This process is repeated until the class distribution is balanced.

SMOTE is visualised in Figure 2.5

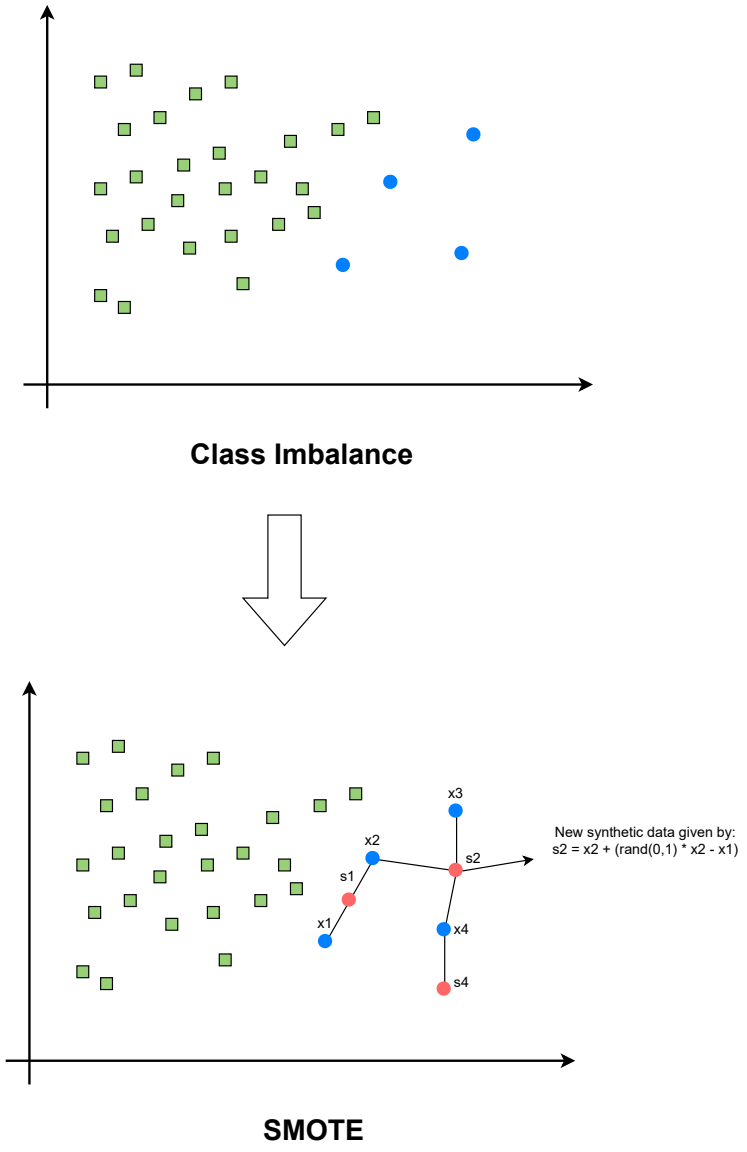

FIGURE 2.5: Depiction of SMOTE [Chawla et al., 2002]

Although SMOTE is effective in addressing class imbalance, it could also introduce the possibility of loss of information. This technique generates samples solely for the minority class, potentially leading to overlooked information in the majority class. Such a loss could impact the classifier's overall performance, particularly in scenarios involving limited dataset sizes.

2.6 Public Transport Landscape in Focus Countries

In Nairobi, public transport accounts for 36.4% of trips made by commuters. The largest share is taken by those who walk or cycle (48.3%), while private car usage holds the smallest market share at 15.3% [Githui et al., 2009]. Although these statistics seem promising, many commuters are captive to their chosen modes due to factors like transport costs. For instance, some may walk not out of preference but due to the un-affordability of bus tickets. Similarly, the severe

traffic congestion in Nairobi might deter individuals from using personal cars, considering the high fuel costs and vehicle wear and tear

Dar es Salaam, Tanzania's largest city with a population of 2.5 million people as of 2003, sees minibuses dominating its public transport scene. As the city expanded both spatially and in population, the demand for public transport surged, which in turn created the need for increased numbers and operations of minibuses as well as the implementation of the Bus Rapid Transport system [Kanyama, 2004; Krüger et al., 2021]. This evolution underscores the city's dynamic response to its growing transportation needs, highlighting the importance of adaptable infrastructure that is responsive to commuter's needs in rapidly developing urban centers.

In South Africa, there's been a consistent push to transition people from private cars to public transport to mitigate issues like traffic congestion and pollution. However, these efforts haven't been very successful. A study by Bubeck et al. [2014] on the proposed BRT system expansion in Gauteng, South Africa, projected that private car usage would continue to rise. This indicates that the majority of public transport users in these countries use it out of necessity rather than choice. Figure 2.6 shows that the percentage of trips made using passenger vehicles will remain high compared to the use of public transport even with proposed expansions on public transport systems.

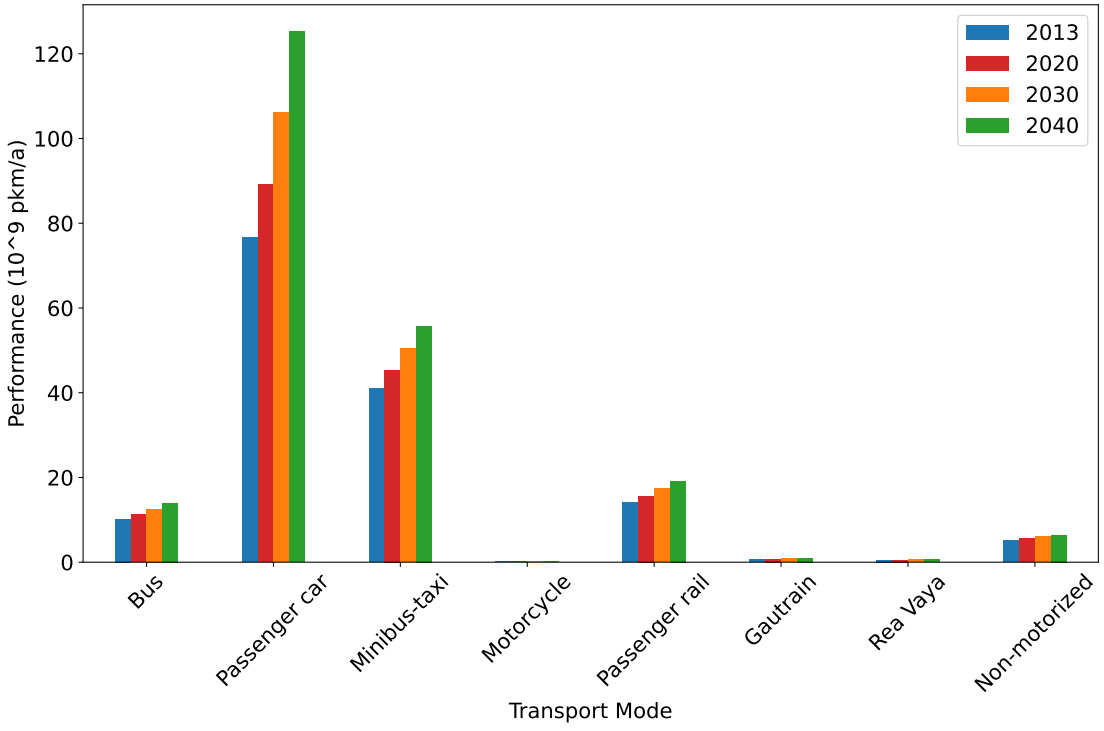

FIGURE 2.6: South African projected public transport mode performance [Bubeck et al., 2014]

The state of public transport in these countries suggests that its use is often not a choice but a circumstance. When given an option, many would prefer private vehicles due to factors like convenience, safety, cost, availability, reliability, and overall service quality [Bubeck et al., 2014; Githui et al., 2009]. This presents a significant challenge for transport engineers and providers aiming to make public transport the preferred choice. This is especially difficult in most sub-Saharan countries where investment in public transport provision is de-prioritised. Given these challenges, a potential solution is to mine commuter sentiment to understand their experiences and preferences, which is the primary objective of this project.

2.7 Discussion

In summarising this chapter, we observe that the use of social media for opinion mining has become an invaluable tool in understanding public sentiment, especially in the context of public transport. Platforms like Twitter and Facebook have revolutionised data collection, offering real-time insights into user experiences and preferences. This shift from traditional data collection methods to digital platforms is not without its challenges, particularly in terms of data privacy and ethical considerations. However, these platforms have proven to be rich sources of user-generated content, essential for capturing a wide range of public opinions.

The integration of social media in public transport systems underscores a growing trend among transit agencies to engage with users more interactively. Campaigns like PRASA's #adoptastation highlight the proactive steps taken by agencies to involve the public in infrastructure maintenance and improvement. However, this also points to a broader issue within the public transport systems of the focus countries of this study, where the majority of users rely on public transport out of necessity rather than choice, often due to the unavailability of alternatives.

The research further identifies a critical gap in NLP resources for low-resource languages. Despite advancements in language models like AfriBERTa, Afro-XLMR, AfroLM, and PuoBERTa, there remains a significant need for inclusive development in language technology, ensuring that advancements in NLP are accessible to all linguistic communities. This is particularly crucial in a multilingual context, where the effectiveness of opinion mining tools can be limited by the lack of resources for low-resource languages.

Overall, this chapter highlights the potential of social media and NLP in enhancing public transport services. By tapping into the wealth of data available on social media platforms and employing advanced NLP techniques, transit agencies and policymakers can gain deeper insights into user experiences. This, in turn, can inform strategies to improve service quality and make public transport a more appealing choice for commuters. However, achieving this requires a balanced approach that considers both the technological aspects of data mining and the socio-cultural dynamics of language and communication.

The subsequent chapter delves into the data processing and EDA conducted on the extracted tweets. This includes insights into the languages within the dataset, identified trends, and the results of the feature extraction process. Training data was sourced from previous studies that conducted sentiment analysis on mixed code data and African languages [Muhammad et al.,

2023a; Mabokela and Schlippe, 2022]. These datasets were considered gold standard and were used to train the Language models in the subsequent chapters.

Chapter 3

Methodology

This research was designed as an exploratory study, with a strong emphasis on experimental methods to navigate the complexities of multilingual datasets and the nuances of code-mixed data. This study used data extracted from social media platforms, with the key focus being public transport related content. The data was sourced from three different countries, namely: Kenya, South Africa, and Tanzania. Given the linguistic tapestry of these nations, our dataset was inherently multilingual, encompassing languages such as English, Swahili, Sepedi, and SeTswana, and, therefore, consideration was taken in analysing code-mixed data and choosing machine learning models that were pre-trained on these respective languages.

The methodology for data extraction, processing, and analysis was systematically structured into five pivotal steps: (i) Sourcing the data; (ii) Data pre-processing and Exploratory Data Analysis; and (iii) Training data description and analysis;

3.1 Sourcing the data

The researcher sourced data from Twitter under guidance of the Data Science for Social Impact research group, focusing on keywords pertinent to public transport within each country. The keywords, as illustrated in Table 3.1, were tailored to capture the unique transport lexicon of each country.

Table 3.1: Transport keywords for each country

Country	**Transport keywords**
Kenya	'matatu', 'kencom', 'citihoppa', 'public transport', 'boda boda', 'bus', 'kbs', 'mathree', 'ma3', 'tuk tuk', 'BasiGo', 'hoppaciti'
Tanzania	'daladala', 'boda boda', 'dart', 'bus', 'train', 'ferry', 'bajaj'
South Africa	'prasa', 'metrorail', 'taxi', 'public transport', 'putco', 'gautrain', 'rea vaya', 'a re yeng', 'bus', 'train', 'shosholoza meyl', 'metrobus'

To ensure relevance, the data extraction was confined to major metropolitan areas in each country: Nairobi, Dar es Salaam, and Johannesburg. These cities, being the nerve centers of public transport, provided a comprehensive view of commuter sentiments. The timeframe for

this data collection spanned from 1^{st} January 2018 to 1^{st} March 2023. The total size of the extracted dataset is presented in Table 3.2.

Country	**Keywords**	**No. of Sentences**
Kenya	boda boda	1 001
	bus	1 001
	matatu	1 001
	public transport	919
	kbs	483
	tuk tuk	416
	ma3	387
	kencom	330
	mathree	213
	hoppaciti	64
	BasiGo	63
	citihoppa	9
Subtotal		**5 887**
South Africa	bus	1 001
	gautrain	1 001
	prasa	1 001
	public transport	1 001
	taxi	1 001
	train	1 001
	rea vaya	528
	putco	474
	metrorail	366
	metrobus	99
	shosholoza meyl	96
	a re yeng	92
Subtotal		**7 661**
Tanzania	bus	909
	daladala	37
	ferry	22
	train	11
	bajaj	8
	dart	7
	boda boda	4
Subtotal		**998**
Total dataset size		**14 546**

TABLE 3.2: Extracted dataset properties

It should be noted that user privacy and ethical data handling were top priority, therefore all personally identifiable information, including usernames and location tags, were systematically removed from the dataset before commencing the analysis. This measure was taken to guarantee complete anonymity of the Twitter users whose data was included in our study.

3.2 Data Processing and Exploratory Data Analysis (EDA)

3.2.1 Overview

The data processing began with meticulous data cleaning, which involved:

- **Punctuation Removal:** Utilising Python's NLTK library, each sentence was tokenized to extract individual words. Subsequently, any punctuation marks, predefined in a comprehensive list, were removed. The list of punctuation includes:

  ```
  ['!"#$%&'()*+,./:;<=>?@[\]^_`{|}~']
  ```

- **Expanding Contractions:** Apostrophe-shortened words were expanded to their full forms. For example, "what's" was transformed to "what is". This modification was applied only to the English words in the dataset.
- **Word Lemmatization:** English words in the dataset were lemmatized using the NLTK library to derive their base or root forms.
- **Stop Word Removal:** Stop words in both English and Swahili were eliminated from the sentences. For English, the predefined stop word list from the NLTK library was employed. For Swahili, a custom list based on domain knowledge was developed and utilised for removal.
- **Named Entity Removal:** Using the NLTK library, named entities classified as Geo-Political Entities, Locations, or Organizations were identified and removed from the dataset.
- **Trimming Spaces:** All sentences were stripped of leading and trailing spaces for cleaner data presentation.
- **Keyword Removal:** Transport-related keywords, initially used in data collection, were removed from the dataset.

The overarching goal of these steps was to highlight the primary features within the data and to delve into its semantic nuances. Throughout the exploration phase, various configurations of this process were tested, aiming to discern the optimal combination that would reveal the richest characteristics of the data.

To identify the languages present in our dataset, we employed tools like *AfroLID* [Adebara et al., 2022], *Franc*, and *CLD3* [Salcianu et al., 2018]. The choice of these tools was influenced by the findings of Adebara et al. [2022], which highlighted their superior performance in detecting African languages. Language Identification (LID) was executed both before and after the data

cleaning, and the outcomes were compared. To ensure the accuracy of our LID results, we undertook a manual validation, sampling each detected language and finalising the language categorisation for the various datasets.

Trend analysis involved examining the volume of tweets associated with the keywords listed in Table 3.1, with particular attention to peaks within the dataset for valuable insights. Following this, feature extraction was conducted, a crucial step in understanding the semantic interplay among words, especially in mixed code datasets. This process not only helped discern potential similarities across different datasets but also provided a deeper understanding of the underlying patterns and themes in the data.

3.2.2 Language Identification

Language identification (LID) played a pivotal role in determining the linguistic diversity within our raw dataset. We employed three distinct tools for this task: *AfroLID*, *Franc*, and *CLD3*.

The raw data was input into each model, and the outcomes are depicted in Figures 3.1 to 3.3. Unfortunately, CLD3 encountered difficulties in recognising Kenyan and South African languages. This issue primarily stemmed from installation constraints, as the necessary dependencies for running the model could not be installed on the local machine used for analysis. Consequently, the results for the Kenyan and South African datasets are exclusively based on AfroLID and Franc, as CLD3's data is not represented due to these technical challenges.

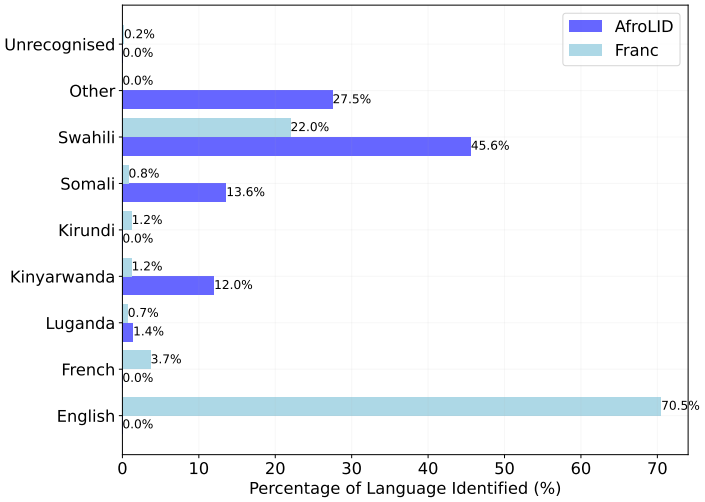

FIGURE 3.1: Language percentage distribution in the Kenyan dataset

Considering the study's focus on English, isiZulu, SeTswana, and Swahili, a manual validation was performed on a sample of 100 entries for each language from every country's dataset. The LID tool demonstrating the highest accuracy and reliability (i.e., consistently producing results) was selected for further tasks. This process was carried out both before and after data cleaning (as detailed in Section 3.2). Notably, the results from the pre-cleaning phase showed superior accuracy, leading to the decision to utilise the pre-cleaned datasets for further LID analysis. Table 3.3 provides a detailed performance comparison of each model based on this manual validation.

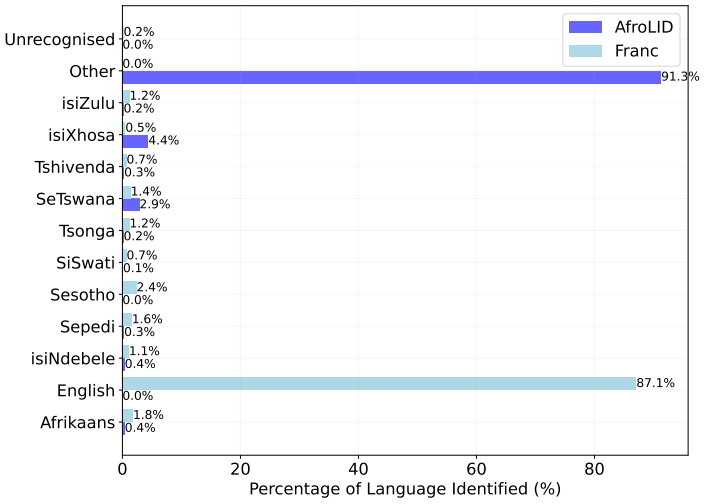

FIGURE 3.2: Language percentage distribution in the South African dataset

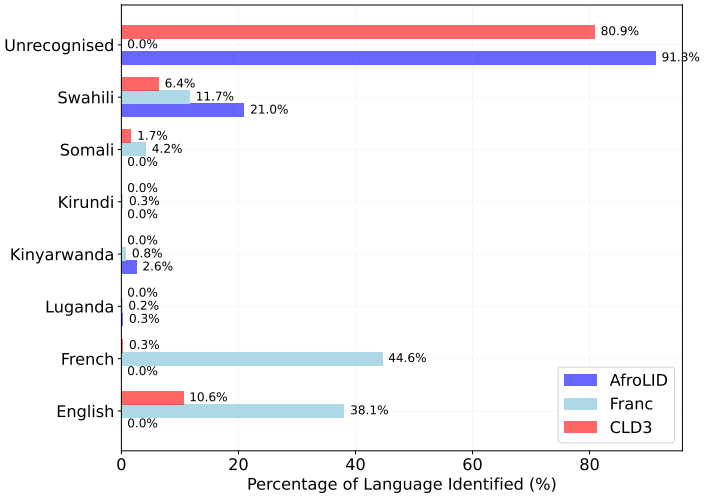

FIGURE 3.3: Language percentage distribution in the Tanzanian dataset

TABLE 3.3: LID validation results

Language	**AfroLID**	**CLD3**	**Franc**
Swahili	48%	84%	73%
SeTswana	50%	-	75%
isiZulu	33%	-	44%
English	-	60%	46%

On average, Franc outperformed CLD3 and AfroLID in the validation process, proving to be the most reliable in terms of result consistency. For example, although AfroLID excelled in identifying African languages, it struggled with English tweets, missing the nuanced linguistic diversity within the data. Additionally, CLD3 encountered difficulties in recognising Kenyan and South African languages due to technical challenges. Based on these findings, Franc was selected as the optimal LID tool for the next stages of this project.

The next step involved identifying code-mixed sentences within the dataset. Since the Language Identification (LID) process was initially performed at the sentence level, and each model predicted the most likely language based on probabilities, it wasn't possible to detect code-mixed

sentences directly. To overcome this, sentences were tokenized and LID was applied on a word-by-word basis. The languages identified within each sentence were then cataloged to ascertain the extent of code-mixing. This detailed analysis was limited to the focus languages of the study: English, isiZulu, SeTswana, and Swahili. We first narrowed down the dataset using the initial LID results from the Franc model, selecting sentences identified as one of the focus languages. This filtration step adjusted the dataset to the size outlined in Table 3.4. Following this, we applied the word-by-word LID method to pinpoint code-mixed sentences, with the findings presented in Table 3.5 and Figure 3.4.

TABLE 3.4: Focus languages dataset size after filtering using Franc results

Language	No. of Sentences
English	11 204
isiZulu	95
SeTswana	111
Swahili	1 407
Total	**12 817**

TABLE 3.5: Dataset size after code mixing language identification

Language	No. of Sentences
Code-mixed (eng-swah)	3 703
Code-mixed (eng-tsn)	1 943
Code-mixed (eng-zul)	608
English	4 777
isiZulu	40
Setswana	970
Swahili	663
Unrecognised	113
Total	**12 817**

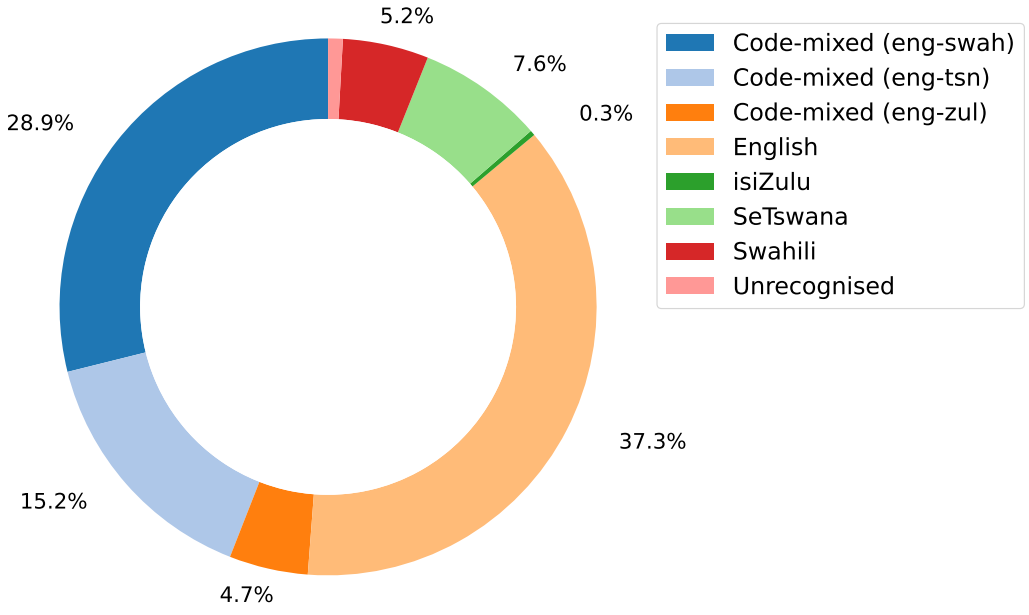

FIGURE 3.4: Focus language distribution of dataset taking into consideration code mixing

The analysis reveals that approximately 50% of the dataset comprises code-mixed content. This significant proportion underscores the importance of factoring in code-mixing when undertaking NLP tasks, particularly for data sourced from social media platforms.

3.2.3 Trend Analysis

Trend analysis was an integral component of our data analysis, aiming to identify possible significant events based on the volume of tweets on specific dates. This exercise was particularly focused on tweets originating from South Africa and Kenya, with the results illustrated in Figures 3.5 and 3.6. Noteworthy events corresponding to these trends are cataloged in Table 3.6. The Tanzanian dataset was excluded from this analysis due to its limited size, which rendered it unsuitable for a comprehensive time series representation.

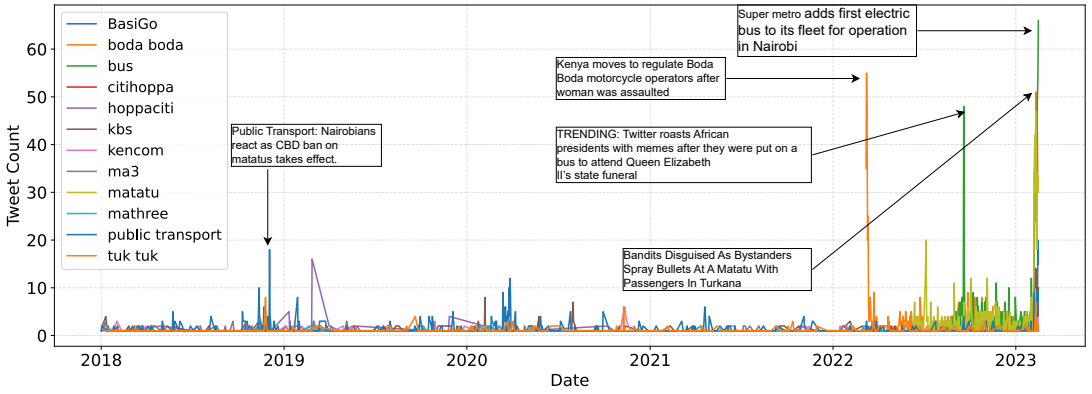

FIGURE 3.5: Trend analysis on Kenyan tweets

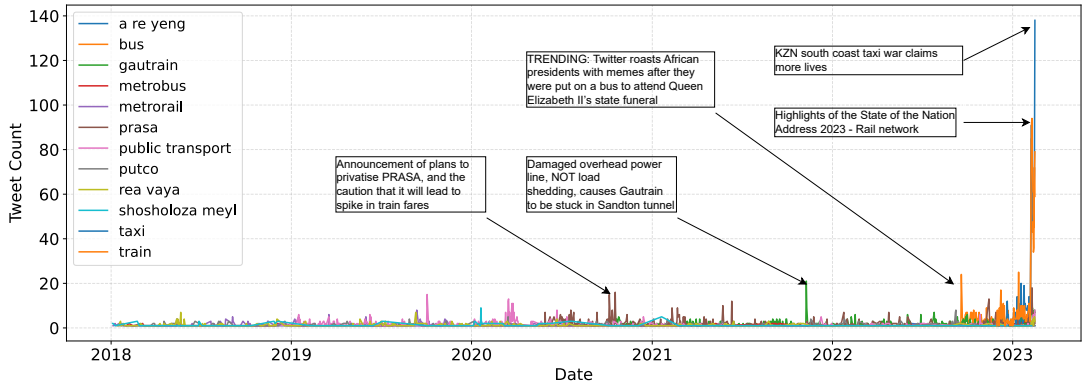

FIGURE 3.6: Trend analysis on South African tweets

The insights from trend analysis show a clear pattern where increased tweet activity often aligns with negative events, though there are notable exceptions like tweets for awareness or information dissemination. In Kenya and South Africa, surges in tweeting commonly followed criminal incidents. Interestingly, both countries shared a distinctive topic, highlighting African presidents using buses to attend Queen Elizabeth's funeral, which reflected the perception of bus travel as inferior and illuminated wider sentiments about public transport in the continent. Conversely, there were also positive instances where tweets acted as informative sources, announcing new

TABLE 3.6: Trend analysis key events

Country	Event Date	Keyword	URL Link
Kenya	2023-02-11	matatu	Bandits Disguised As Bystanders Spray Bullets At A Matatu With Passengers In Turkana
	2018-12-03	public transport	Public Transport: Nairobians react as CBD ban on matatus takes effect
	2023-02-15	bus	Super metro adds first electric bus to its fleet for operation in Nairobi
	2022-03-09	boda boda	Kenya moves to regulate Boda Boda motorcycle operators after woman was assaulted
	2022-09-19	bus	TRENDING: Twitter roasts African presidents with memes after they were put on a bus to attend Queen Elizabeth II's state funeral
South Africa	2023-02-15	taxi	KZN south coast taxi war claims more lives
	2023-02-09	train	Highlights of the State of the Nation Address 2023 - Rail network
	2020-10-18	prasa	Announcement of plans to privatise PRASA, and the caution that it will lead to spike in train fares
	2021-11-09	gautrain	Damaged overhead power line, NOT load shedding, causes Gautrain to be stuck in Sandton tunnel
	2022-09-19	bus	TRENDING: Twitter roasts African presidents with memes after they were put on a bus to attend Queen Elizabeth II's state funeral

developments like the introduction of electric buses or significant mentions of public transport improvements in state of the nation addresses.

3.2.4 Feature Extraction

In the field of Natural Language Processing (NLP), feature extraction is a pivotal step that involves transforming raw data into a set of features or attributes that can be more easily interpreted and analyzed. Essentially, it's about distilling the essence of the data into a format that's more amenable to machine learning or other analytical techniques. In the context of text data, this often means identifying patterns, relationships, or characteristics that can help in understanding the underlying semantics or sentiment of the content.

For this study, feature extraction was conducted by determining word embeddings derived using the Word2Vec model. Despite Word2Vec being a non-contextual embedding, it effectively captures syntactic and semantic word similarities. We decided to used this model because its architecture offered a balance between simplicity and the ability to capture rich word relationships. Word embeddings are vector representations of words capturing their semantic meanings. Subsequent clustering (K-Means clustering) was performed to discern the relationships between these words. This was instrumental in probing the semantic connections between different words, especially in datasets with mixed code data. Each country's dataset underwent this process, and insights were derived from the outcomes. The results revealed that topic modeling could be

applied to each dataset, and themes could be inferred from the clusters of keywords and the primary features extracted from the data.

The first step in the feature extraction process was to perform a straightforward word count analysis. This helped ascertain the most prevalent words within each dataset, laying the groundwork for potential theme identification. This analysis was carried out after data processing. Figures 3.7 to 3.9 present the word counts derived from each dataset.

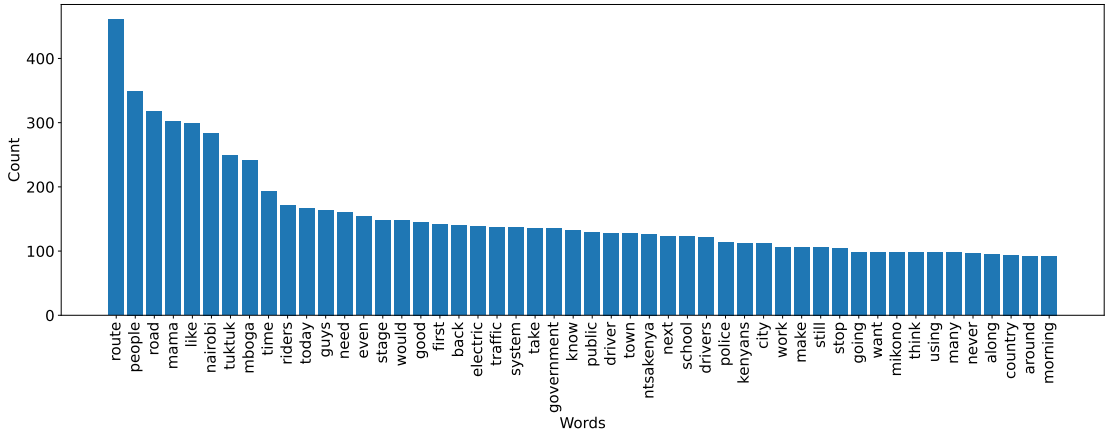

FIGURE 3.7: Kenyan word count

Within the Kenyan dataset, transport-related terms such as "route", "road", "tuktuk", "stage", "traffic" emerged as dominant. However, other terms not directly tied to transport, such as "government", "school", "police", also made a significant appearance. This underscores the interconnectedness of these entities in the public discourse around transport. For instance, one of the government's functions is the provision of public transport, therefore the conversation surrounding public transport cannot occur without government involvement. Also, the term "school" suggests the influence of school schedules on traffic patterns. The presence of "police" alludes to their role in public transport, either in law enforcement or, in some cases, in corrupt practices.

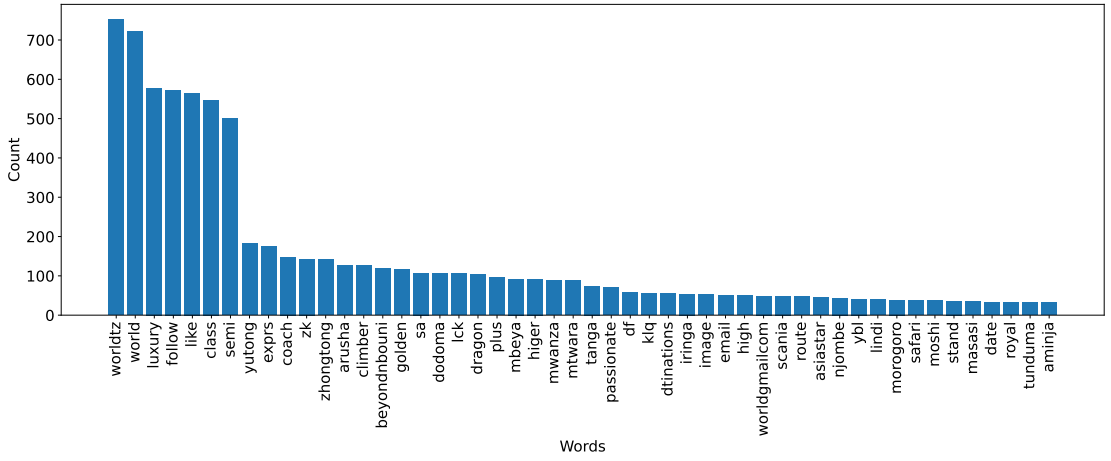

FIGURE 3.8: Tanzanian word count

The Tanzanian dataset paints a unique picture when it comes to word distribution. Predominantly, the dataset is dominated with terms that hint at the promotion of public transport services. Words like “luxury”, “follow” (a nod to its Twitter origin), and “class” suggest a strong advertising undertone. This pattern implies that in Tanzania, transport-related tweets are often crafted by transport providers keen on marketing their offerings. Following this, geographic destinations such as “arusha”, “dodoma”, “mbeya”, “mwanza”, “tanga”, and “morogoro” emerge as recurrent themes. Such a trend underscores the notion that Tanzanian tweets, in the context of transport, lean more towards service promotion and route information rather than capturing raw commuter sentiment. This could pose challenges for those aiming to extract genuine passenger feedback, as the platform appears to be a hub for advertisements and route updates.

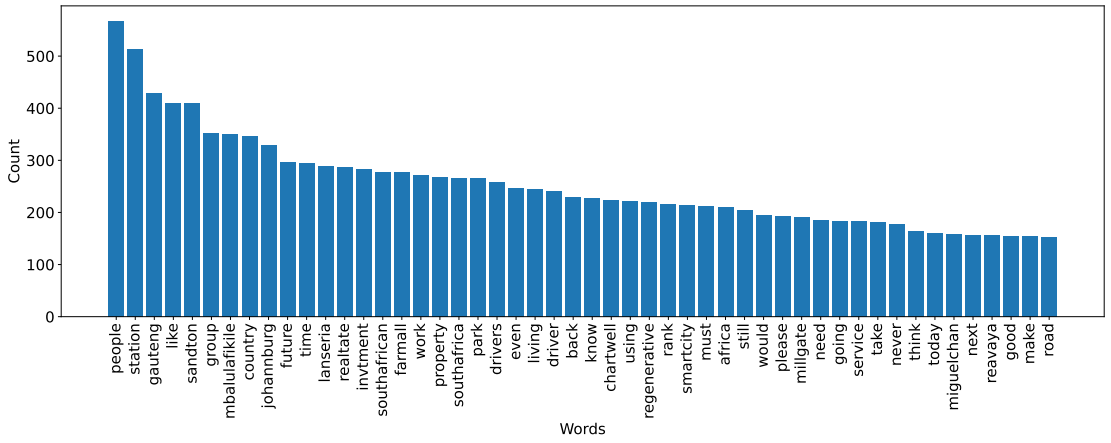

FIGURE 3.9: South African word count

In the South African context, the word count revealed a blend of transport-related terms and those indirectly tied to transport. For instance, terms like “station”, “go”, “park”, “drivers”, “time” can be associated with transport activities, while terms like “mbalulafikile”, referencing the transport minister at the time, hint at government involvement. Similarly, destination-related terms like “gauteng”, “sandton”, “johannesburg”, and “chartwell” were prevalent. This suggests that South African transport-related tweets serve dual purposes: disseminating information and mining public opinion.

To further investigate the intricacies of the data and understand the interplay between various features, clustering was employed on the extracted word embeddings. This technique highlighted the semantic relationships that existed within the datasets. The results, showcasing the principal features from each dataset, are presented in Figures 3.10, 3.11, and 3.12. The clustering was structured based on the number of keywords that informed the data collection for each country. For instance, the Kenyan dataset was segmented into 12 clusters, mirroring the 12 keywords that guided its sourcing (as referenced in Table 3.1). The figures also include the text labels of the centroids within each cluster, offering a glimpse into the primary features within each segment.

In the context of Kenya, the primary feature of “kshs” was associated with a range of topics related to advertising products and public transport pricing, such as bus tickets. The feature “mini” had a strong connection with activities related to mini-buses, while also often mentioning the activities of the minister, possibly indicating its derivation from the word ”minister.” Moreover, it encompassed posts about the COVID-19 pandemic, urging citizens to minimize close contact

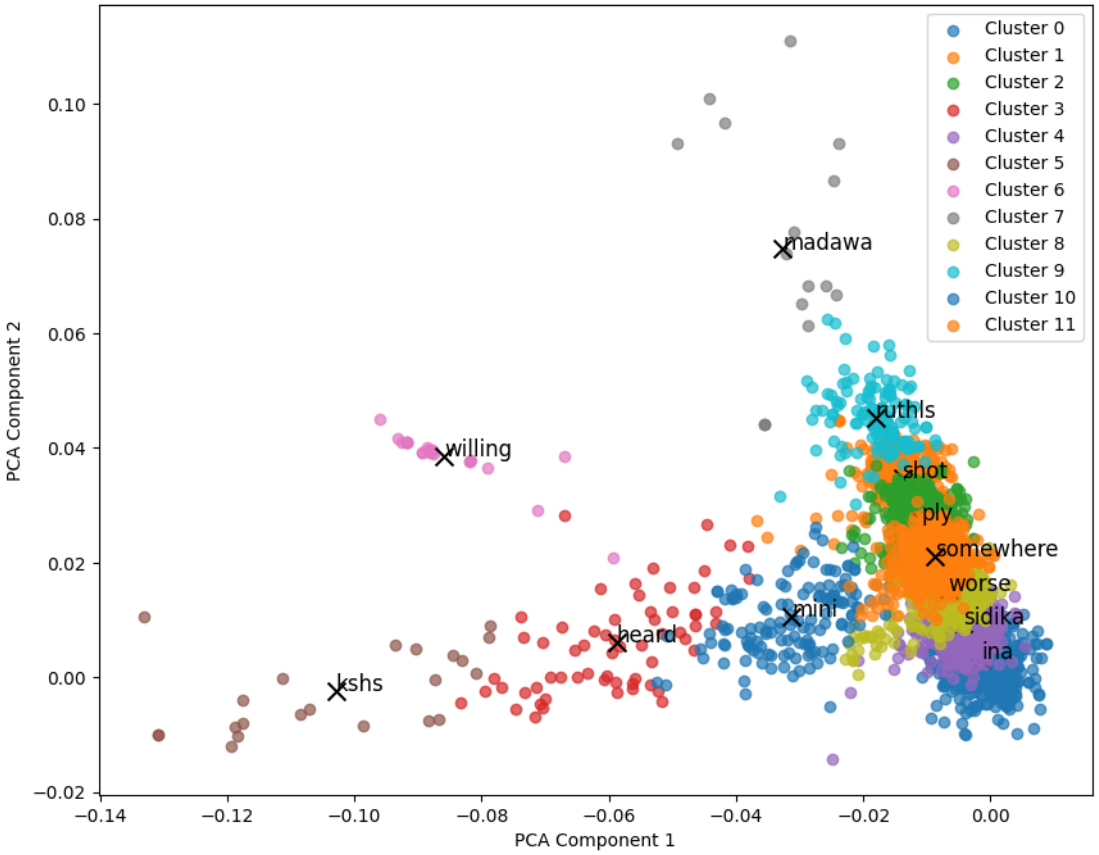

FIGURE 3.10: Main features extracted from the Kenyan dataset

and large gatherings. The "somewhere" feature was closely linked to "kbconglomerate," likely associated with Kenya Power, and commonly related to public transport experiences in specific areas, serving as a sentiment indicator. The term "willing" was prevalent in tweets focused on advertising products and potential job opportunities. Conversely, "worse" referred to sentiment-driven tweets reflecting negative experiences of public transport, discussing issues like drug usage and government-related incidents. "Ruthls" was related to tweets referring to police, military, cartels, and government entities, revealing the sentiment towards these topics. "Madawa", translating to "drugs" in Swahili, referred to a criminal trend of drugging passengers to steal from them, connected to *matatu* travel and providing insights into safety sentiment. "Heard" entailed tweets that retold incidents, spanning COVID-19-related deaths, government activities, and daily social events. These tweets often acted as engagement and information dissemination posts. The term "ina" in Swahili implied "has" and frequently appeared in engagement-type tweets with no dominant theme. "Shot" encompassed tweets describing incidents involving gunshots, highlighting the link between violent crime and public transport, thus revealing negative sentiment. "Sidika" referred to socialite Vera Sidika's travel using Bishop concierge services, indicating the availability of different transport services for Kenyans. "Ply" remained unclear in categorization and carried a predominantly negative sentiment.

Shifting focus to Tanzania, the "brt" feature predominantly revolved around the Bus Rapid Transit (BRT) system in Dar es Salaam. Features like "temeke" spotlighted the Temeke district, while "ukwelidaima" resonated with trending government-centric hashtags. "Umeongezeka,"

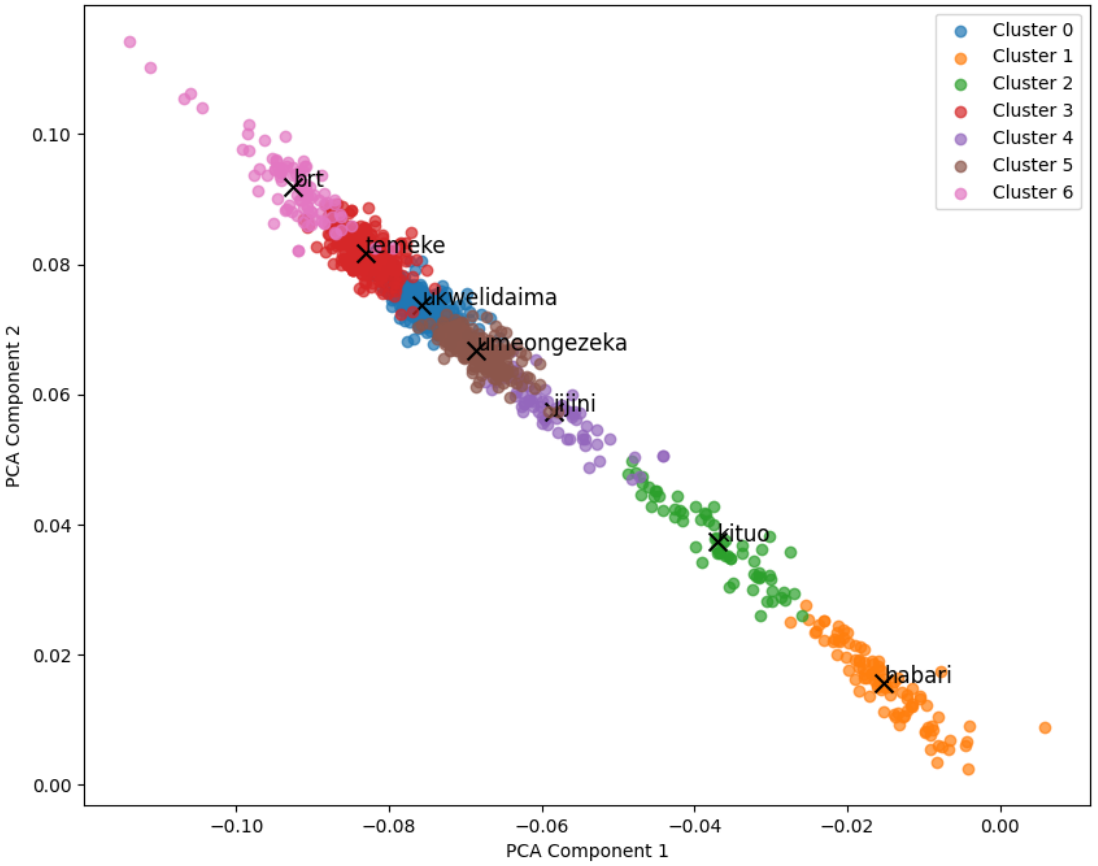

FIGURE 3.11: Main features extracted from the Tanzanian dataset

translating to "has increased" in Swahili, possibly hinted at fare hikes. Features like "Jijini" and "kituo" brought the city and station into focus, respectively, while "habari" echoed news dissemination

In the South African landscape, the "group" feature predominantly revolved around the PRASA group (Passenger Rail Agency of South Africa) and its challenges. Features like "newpietretrief" spotlighted specific destinations, while "Destroyed" and "failed" echoed the deteriorating state of the public transport system. "Vandalism" encapsulated narratives around damage to the transport infrastructure, particularly the rail system. "Tumisole" possibly alluded to a user named Tumi Sole, while "titomboweni" resonated with discussions around Tito Mboweni, the then Finance Minister, and his association with PRASA's challenges. "Today" captured daily happenings, often emphasizing negative transport incidents, while "manhlamza" remained ambiguous in its context.

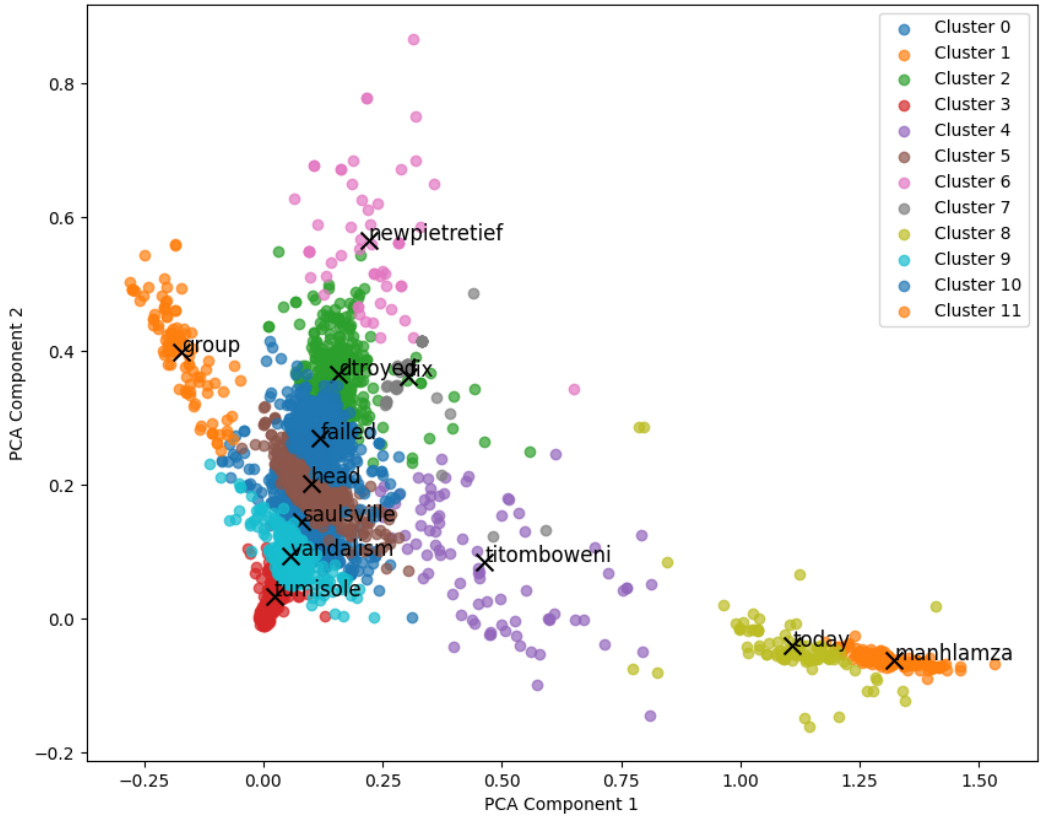

FIGURE 3.12: Main features extracted from the South African dataset

3.3 Training Dataset Description

The properties of the training datasets used in the study are summarised in Table 3.7

TABLE 3.7: Training datasets

Dataset	**Sources**	**Languages**	**Quality**	**Labels**
Afrisenti	Muhammad et al. [2023a]	Swahili	High	Positive, Neutral, Negative
DSFSI SeTswana	DSFSI	SeTswana	Medium	Positive, Neutral, Negative
DSFSI isiZulu	DSFSI	isiZulu	Medium	Positive, Neutral, Negative
DSFSI code-mixed	Researcher (DSFSI)	English, Swahili, isiZulu, SeTswana	Low	Positive Neutral, Negative

Further details of the training datasets are presented in the subsequent subsections.

3.3.1 *AfriSenti* Swahili training datasets

The study employed the *AfriSenti* Swahili training datasets [Muhammad et al., 2023a] as the foundation for the training phase. This dataset is a compilation of tweets labeled to reflect the three traditional sentiment categories: positive, neutral, and negative. A visual representation of the label distribution within this dataset is presented in Figure 3.13.

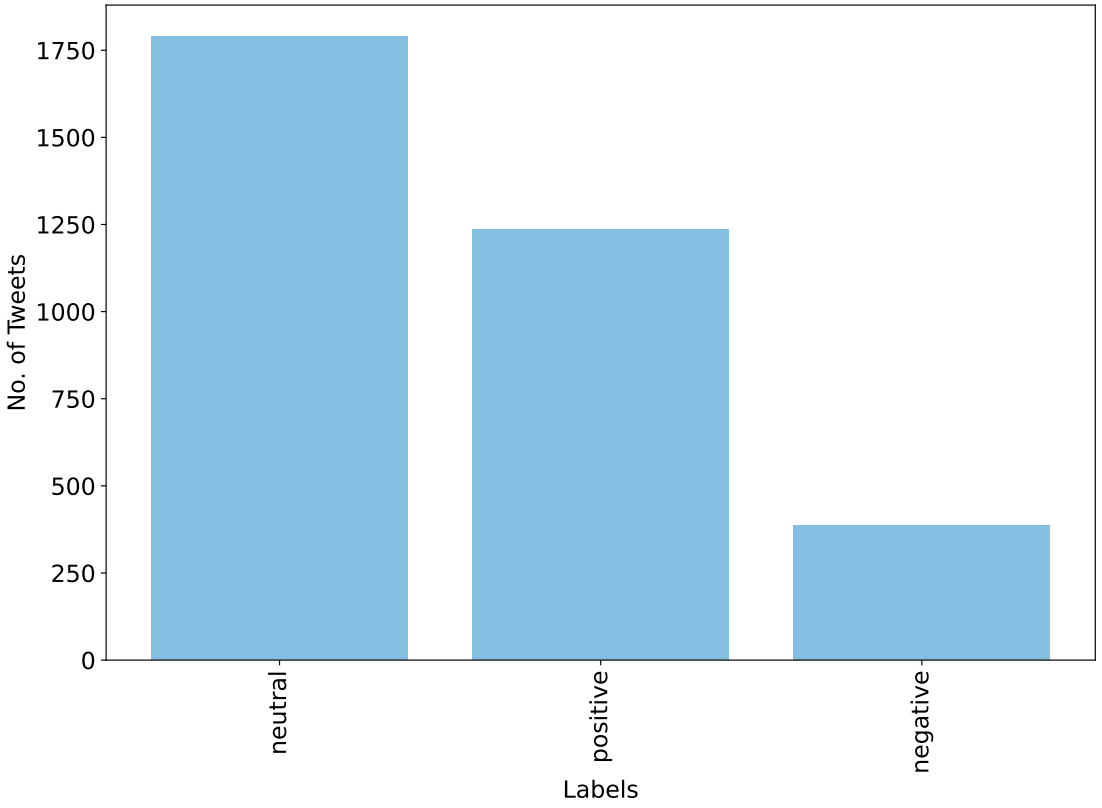

FIGURE 3.13: Swahili training dataset label distribution

A significant observation from the Swahili training dataset is the dominance of neutrally labeled tweets. This skewness in the labeling may necessitate potential data adjustments during the training process. This would be done to ensure a balanced representation of labels, which in turn would aid in minimiSing biases in the resultant model. To gain deeper insights into the dataset's content, an analysis was conducted on the frequency of words within it. The findings from this analysis are illustrated in Figure 3.14.

Delving into the most recurrent words in the Swahili dataset provided a window into the prevailing sentiment. Words such as "changamoto" and "tatizo", translating to "challenge" and "problem" respectively, hint at potential areas of concern or discussion points. Additionally, terms such as "serikal" and "rais", which translate to "government" and "president", resonate with themes from the research dataset, suggesting political or governance-related discussions. Moreover, the presence of words like "habari", "taarifa", and "mambo", all pointing towards

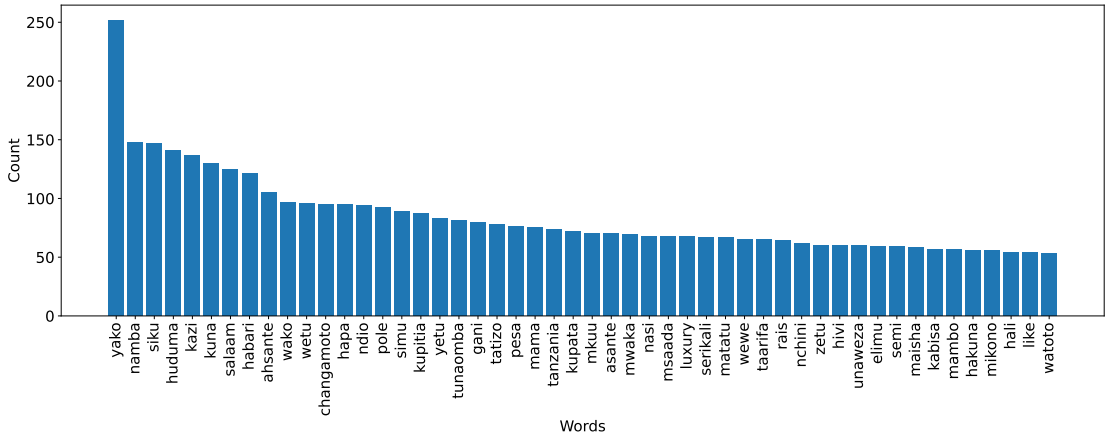

FIGURE 3.14: Swahili training dataset common word count

information dissemination, further aligned with the overarching themes of the research dataset, emphasising the significance of communication and information sharing in the dataset's context.

3.3.2 DSFSI Setswana and isiZulu training datasets

The annotated datasets provided by the Data Science for Social Impact (DSFSI) were meticulously compiled, focusing on tweets selected based on language identification. This method ensured that the extracted tweets were relevant to specific linguistic demographics. The datasets underwent a detailed manual annotation process to categorise them into the three conventional sentiment classes: positive, neutral, and negative. This task was performed by 2 to 3 annotators per tweet, with the final sentiment label being assigned based on the majority agreement among the annotators. The label distributions for these annotated datasets are visually represented in Figures 3.15 for Setswana and 3.16 for isiZulu. The SeTswana dataset consisted of 168 validly annotated tweets, while the isiZulu dataset contained 180 annotated tweets.

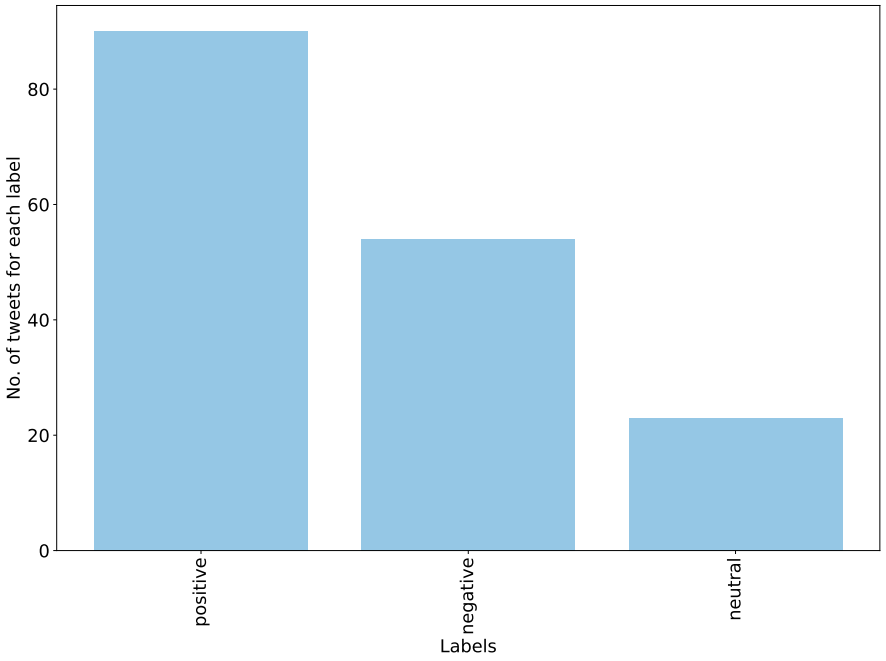

FIGURE 3.15: SeTswana dataset label distribution

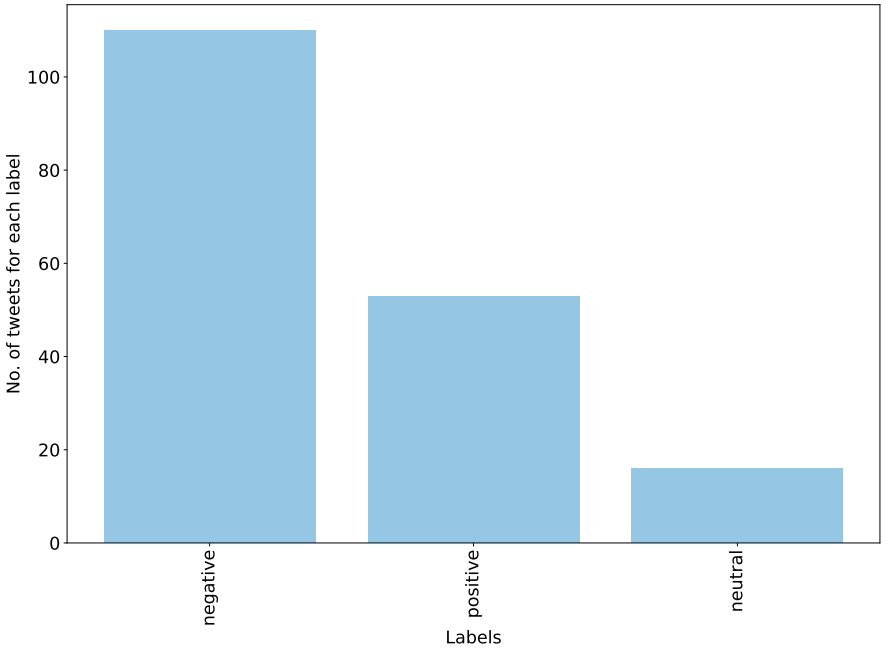

FIGURE 3.16: isiZulu dataset label distribution

3.3.3 DSFSI Code-Mixed Dataset Creation

The idea for creating a code-mixed dataset stemmed from the observation that social media data, especially in the countries included in our study, often comprises code-mixed content. It is a prevalent practice for individuals to blend English with African languages when expressing opinions on social media platforms. In our dataset, code-mixed data, constituted a significant portion, accounting for 49% of the entire dataset (See Figure 3.4). This prevalence underlined

the necessity of incorporating a code-mixed dataset into our model development and evaluation framework.

The development of this dataset involved initially identifying code-mixed entries within our collected data (refer to the method outlined in Section 3.2.2). Following this, we adopted an auto-labeling approach, utilising emojis present in the tweets as indicators of sentiment. This method drew inspiration from the paper by [Choudhary et al., 2018]. The final dataset comprised 881 tweets, with the label distribution illustrated in Figure 3.17.

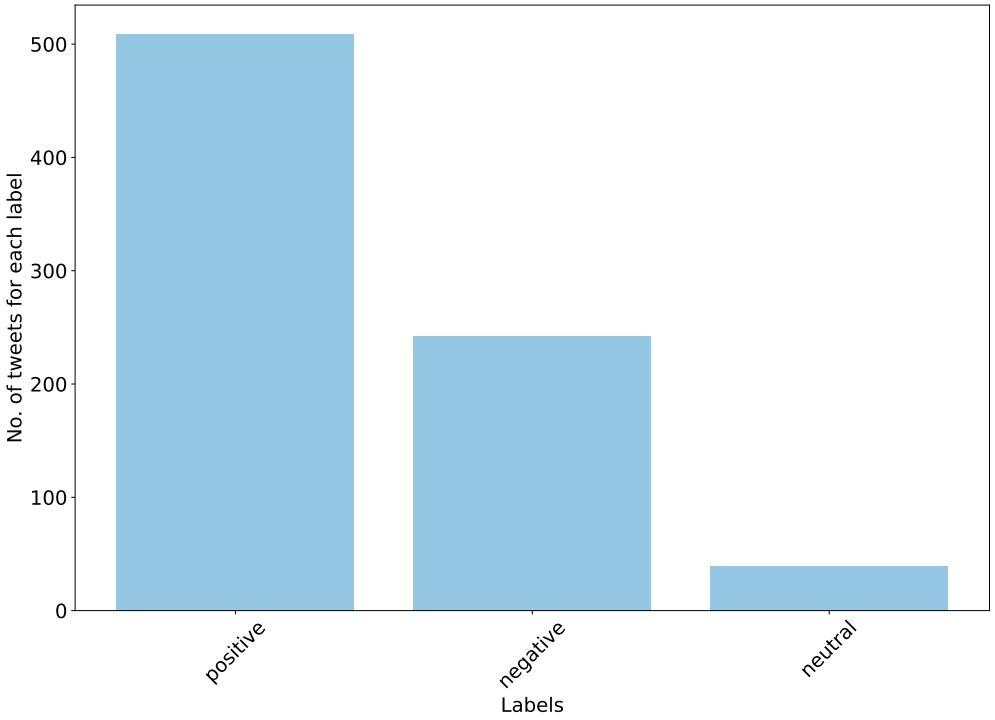

FIGURE 3.17: Label Distribution of the Code-Mixed Dataset

However, this auto-labeling process necessitates thorough validation. We recommend conducting manual validation, aiming for a True Positive Rate (TPR) of at least 0.7 to affirm the dataset's validity. A notable issue revealed in the label distribution is the severe imbalance, with a skew towards positive labels. Closer examination indicated that relying solely on emojis for auto-labeling can be misleading. For instance, positive emojis are often employed ironically or comically, which could distort the true sentiment of a message. Therefore, future manual validation is essential to ensure accuracy.

While auto-labeling is beneficial in scenarios with limited annotator availability, it demands rigorous manual validation. Consequently, there is a trade-off to consider when devising an annotation strategy, balancing the efficiency of auto-labeling with the accuracy of manual validation.

3.3.4 Handling Imbalanced Datasets

Figure 3.18 to 3.21 present the class distribution within the different training datasets after the application of SMOTE. This process involved tokennizing the tweets within the datasets and

extracting the embeddings using a Word2Vec model. The word embeddings were then used to generate the synthetic sample. The process is outlined in Section 2.5.

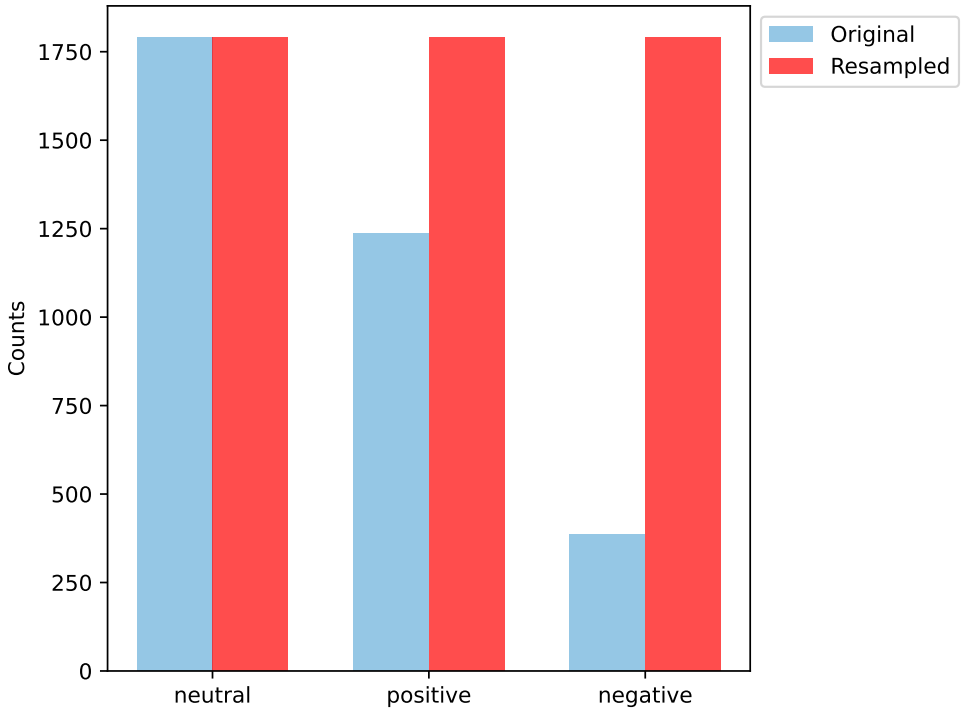

FIGURE 3.18: Label Distribution of Swahili Dataset after the application of SMOTE

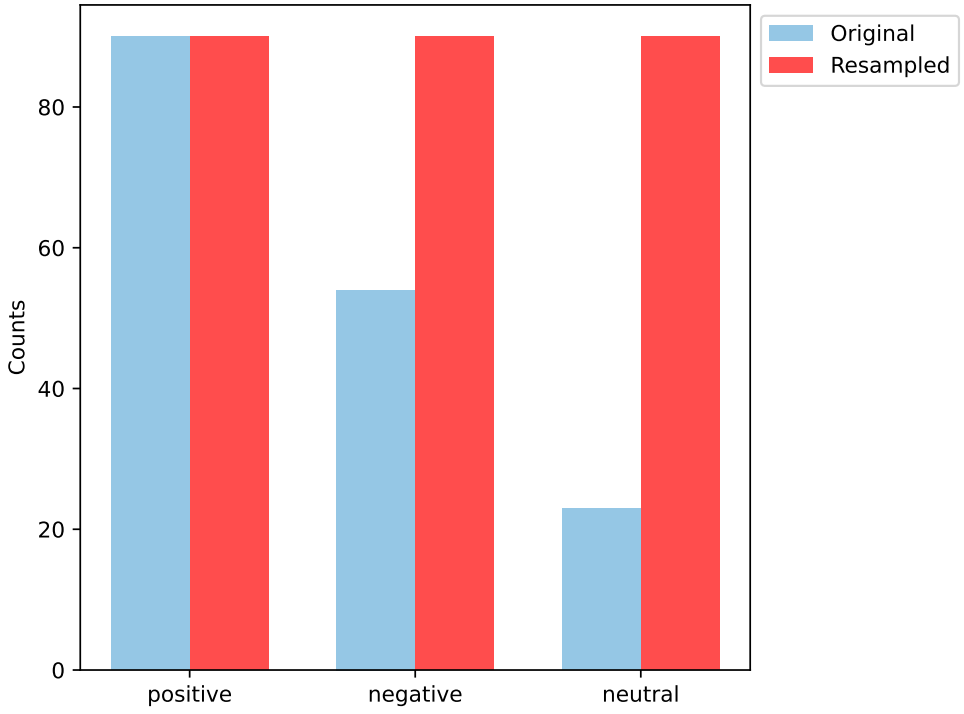

FIGURE 3.19: Label Distribution of SeTswana Dataset after the application of SMOTE

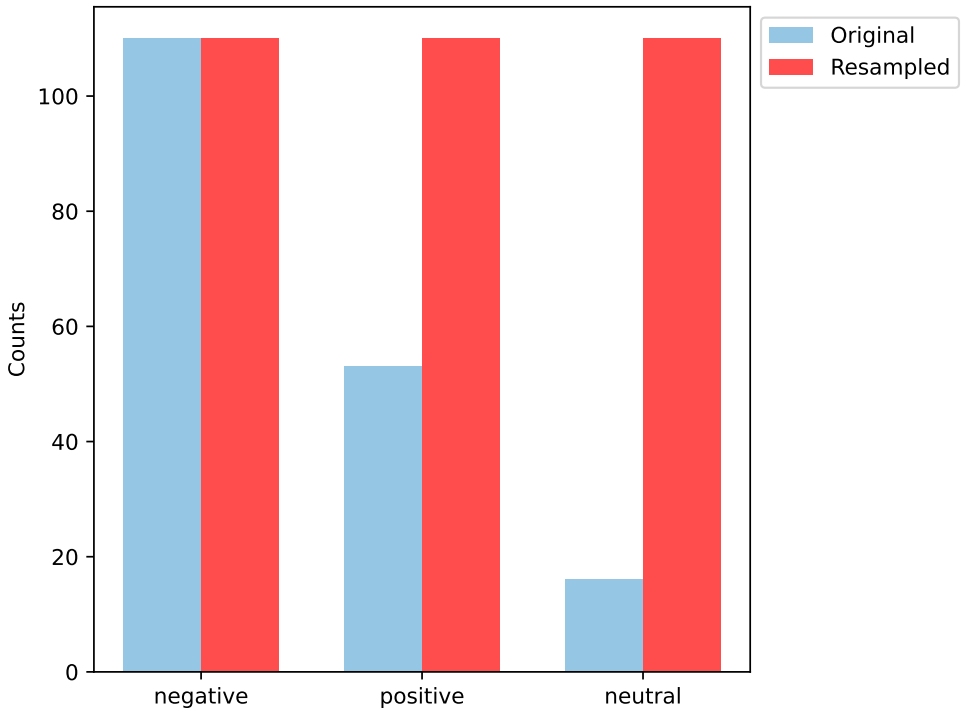

FIGURE 3.20: Label Distribution of isiZulu Dataset after the application of SMOTE

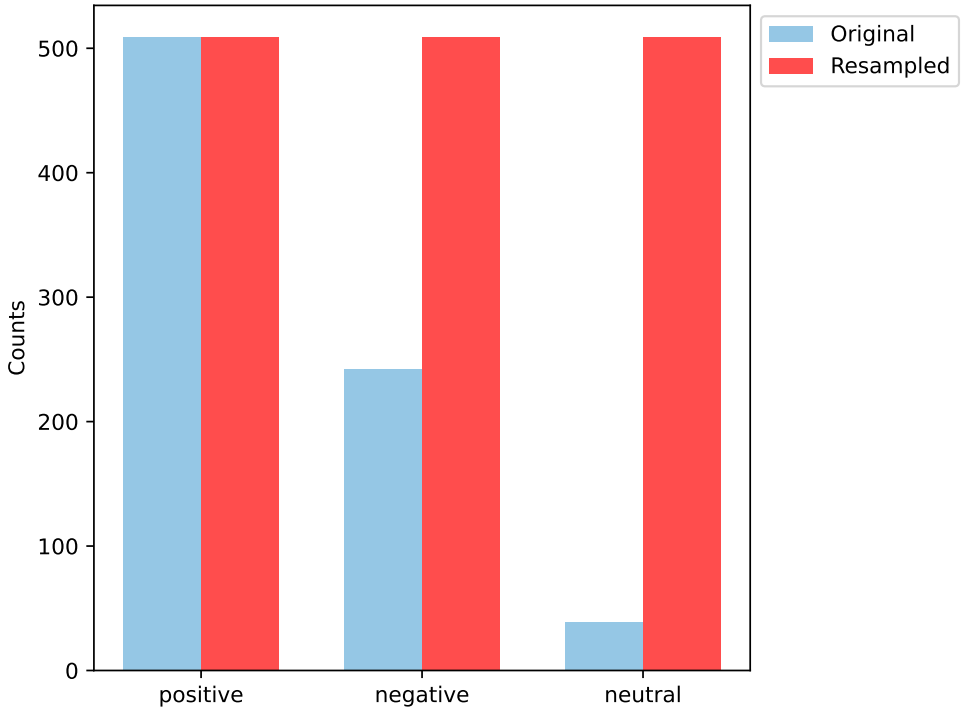

FIGURE 3.21: Label Distribution of Code-mixed Dataset after the application of SMOTE

3.4 Discussion

The exploration of various datasets and methodologies in this study provided a comprehensive understanding of the sentiment analysis landscape, particularly in the African context. Starting with the Language Identification (LID) task, the study revealed the dominance of English and Swahili in the datasets. This dominance underscored the significance of low resource languages such as Swahili in the African Twitter landscape, emphasising the need for specialised models tailored to these languages to ensure accurate sentiment analysis.

The trend analysis further enriched our understanding by highlighting the impact of key events on Twitter activity. The analysis, particularly of the South African and Kenyan tweets, showcased how specific incidents or occurrences can significantly influence the volume and sentiment of tweets. This observation is crucial as it underscores the reactive nature of Twitter data, which can be leveraged for real-time sentiment analysis.

The feature extraction process, which delved into the semantic relationships between words, was instrumental in understanding the underlying themes within the datasets. The use of word embeddings derived from the Word2Vec model and clustering (K-Means clustering) techniques illuminated the intricate relationships between various terms, especially in mixed code data. This process not only highlighted the primary themes within the datasets but also emphasised the importance of context in sentiment analysis.

Lastly, we delved into an extensive exploration of various training datasets, including the *AfriSenti* Swahili training dataset, the DSFSI Setswana and isiZulu datasets, and the specially crafted Code-mixed dataset. Our examination provided valuable insights into the characteristics of the available labeled data, crucial for effective model training. Notably, the analysis of label distributions and word frequencies in these datasets brought to light key challenges and potential avenues for enhancing sentiment analysis models. We identified issues such as class imbalance and the prevalence of certain recurrent terms, underscoring the necessity for sampling procedures such as SMOTE and thoughtful data modifications. The annotated dataset sizes are summarised in Table 3.8 below:

Table 3.8: Summary of training datasets

Dataset	Language	Train	Dev	Test	**Total**
AfriSenti (swah)	Swahili	1 810	453	748	**3 011**
DSFSI (tsn)	SeTswana	-	-	168	**168**
DSFSI (zul)	isiZulu	-	-	180	**180**
Code-mixed	eng-swah	-	-	430	**430**
	eng-tsn	-	-	345	**345**
	eng-zul	-	-	106	**106**

In conclusion, the findings from this chapter set the stage for the subsequent phases of the research. The next chapter will delve deeper into model development and evaluation. Given the insights from the current exploration, special emphasis will be placed on data modification to ensure a balanced and representative training process.

Chapter 4

Model Development and Evaluation

4.1 Overview

This study's model training and evaluation were conducted using the NVIDIA Tesla T4 GPU. This GPU, based on the Turing architecture, boasts 16 GB of GDDR6 VRAM, providing ample power for handling complex tasks. Our research primarily focused on the application of Pre-trained Language Models (PLMs) and the development of a Siamese Neural Network. The latter was specifically designed to address the code-mixed nature of our dataset. The process of developing the Siamese Neural Network included establishing its architecture, training and evaluation. Detailed insights into these processes are provided in Section 4.3. For the descriptions of all the models utilised in this study, please refer to Section 4.2.

The PLMs employed in our study were *AfriBERTa*, *AfroXLMR*, *AfroLM*, and *PuoBERTa*. These models underwent rigorous testing using annotated data to evaluate their efficacy in sentiment prediction and establish performance benchmarks. Each model was aligned with the language(s) it was initially trained on, and its effectiveness was gauged using the F1-score metric. To further enhance the models' performance, we undertook fine-tuning, data augmentation, and re-training procedures. These procedures are thoroughly discussed in Section 4.3.

Subsequently, we scrutinised the reliability of the models' predictions through a detailed result validation process, as outlined in Section 4.4.1. This phase involved manually re-annotating a subset of the tweets used for sentiment prediction and employing the True Positive Rate (TNR) as our primary validation metric. This validation exercise not only shed light on potential inaccuracies within the models but also assisted in pinpointing the most reliable language model among those tested.

In Section 4.4, we present a comparative analysis of the sentiments predicted by these models against the ratings provided by public transport providers. This comparison aimed to glean insights into user experiences and identify any potential gaps between the service quality perceptions of providers and end-users.

4.2 Model Description

This study incorporated a suite of machine learning models, comprising Pre-trained Language Models (PLMs) and a Siamese Neural Network. Each of the PLMs were selected and paired with language datasets corresponding to the languages they were initially trained on. The PLMs in focus were *AfriBERTa*, *AfroXLMR*, *AfroLM*, and *PuoBERTa*. These models were rigorously evaluated for sentiment classification using datasets detailed in Section 3.3.

The Siamese network was trained on the code-mixed dataset that was created during this study, employing a novel approach that paired text inputs comprising a high-resource language (code-mixed data of English and Swahili) with a low-resource language (AfriSenti Swahili dataset). This strategy was aimed at harnessing the strengths of the high-resource language to enhance the model's performance. The details of the process followed in creating the code-mixed dataset are outlined in Section 3.3.3. A succinct description of each model is provided in Section 2.4, while Table 4.1 offers a comprehensive overview of their respective properties.

Table 4.1: Summary of the model properties

Model	Size	Languages included
AfriBERTa	126M	Afaan Oromoo, Amharic, Gahuza, Hausa, Igbo, Nigerian Pidgin, Somali, **Swahili**, Tigrinya and Yorùbá
AfroXLMR	550M	Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Sesotho, Oromo, Nigerian-Pidgin, Kinyarwanda, Kirundi, Shona, Somali, **Swahili**, isiXhosa, Yoruba, **isiZulu**, Arabic, French, and **English**
AfroLM	264M	Amharic, Afan Oromo, Bambara, Ghomalá, Éwé, Fon, Hausa, Igbo, Kinyarwanda, Lingala, Luganda, Luo, Mooré, Chewa, Naija, Shona, **Swahili**, **SeTswana**, Twi, Wolof, Xhosa, Yorùbá, and **isiZulu**
PuoBERTa	83.5M	**SeTswana**
Siamese Network	0.17M	Not pre-trained

4.3 Model Development, Evaluation, and Experimentation

Our research undertook a comprehensive evaluation of Pre-trained Language Models (PLMs) using the annotated datasets detailed in Section 3.3. This endeavor was conducted alongside the training, testing, and evaluation of the Siamese Network. Initially, we established baseline results for the PLMs. Following this, we embarked on a phase of model fine-tuning to boost their performance. In cases where further enhancements were necessary, we utilised data augmentation techniques and retrained the PLM on these augmented datasets, subsequently evaluating their performance.

The data augmentation methods implemented in this study were categorised into both lexical and context-based techniques. Lexical techniques, namely Synonym Replacement and Random Swap, were employed to modify individual words within sentences while maintaining the overall context and meaning of the sentence. On the other hand, the context-based technique utilised

focused on paraphrasing to introduce varied contexts into the sentences, thereby enriching the dataset. The specific method used for this purpose was Back Translation. A summary of each method is presented below:

- **Synonym replacement**: This technique involved replacing selected words in the annotated dataset with their synonyms. It required the creation of synonym lists, a task particularly challenging for low-resourced languages. For the AfriSenti (swa) dataset, a tailored list of synonyms was developed. The replacement process was executed exclusively for the Swahili dataset, thanks to the availability of a native Swahili speaker who ensured the synonyms were contextually appropriate. This method was also applied to the Code-mixed dataset but was limited to the Swahili and English segments, utilising the English synonym list from the NLTK library.
- **Random Swap**: This method entailed randomly exchanging positions of two words within a sentence. It was uniformly applied across all training datasets, introducing variability in sentence structures.
- **Back Translation**: Here, sentences were translated into a different language and then retranslated back into the original language, effectively creating paraphrased versions. This paraphrasing was carried out using the GoogleTranslator library and was performed on all training datasets, except the Code-mixed dataset.

After augmenting the data, the newly created augmented sentences were reintegrated into the original datasets. This process significantly increased the size of the datasets, in some cases even doubling the amount of data available for analysis. The augmented dataset sizes are detailed in Table 4.2.

Augmentation process	Dataset	Language	Test data size for model evaluation
Synonym Replacement	AfriSenti (swah)	Swahili	1 496
	Code-mixed	eng-swah	860
		eng-tsn	345
		eng-zul	106
Random Swap	AfriSenti (swah)	Swahili	1 496
	DSFSI (tsn)	SeTswana	336
	DSFSI (zul)	isiZulu	360
	Code-mixed	eng-swah	860
		eng-tsn	690
		eng-zul	212
Back Translation	AfriSenti (swah)	Swahili	1 496
	DSFSI (tsn)	SeTswana	336
	DSFSI (zul)	isiZulu	360

TABLE 4.2: Augmented dataset size

Part of the evaluation process for the Pre-trained Language Models (PLMs) involved conducting manual validation of sentiment predictions to calculate the True Positive Rate (TPR). This crucial step was instrumental in assessing the reliability of the model predictions. Additionally, we evaluated other essential metrics such as GPU usage, processing power requirements, and power consumption. These metrics provided valuable insights into the resource demands of each model, which are vital for their practical deployment. The PLMs evaluation process is visually

depicted in Figure 4.1, serving as a graphical representation of our methodology. The primary distinction between fine-tuning and retraining the model lies in the datasets used. Fine-tuning was conducted using the original annotated datasets, whereas retraining employed the augmented datasets.

Regarding the Siamese Network, our primary objective was to ascertain its effectiveness in managing the code-mixed nature of our dataset. This endeavor entailed developing the model architecture and training it on our available dataset to establish benchmarks crucial for determining its real-world applicability. The methodology adopted for the development of the Siamese Network is detailed in Figure 4.2. To maintain consistency in our research, the Train, Test, and Validation datasets for each language were kept constant throughout the study (refer to Table 3.7).

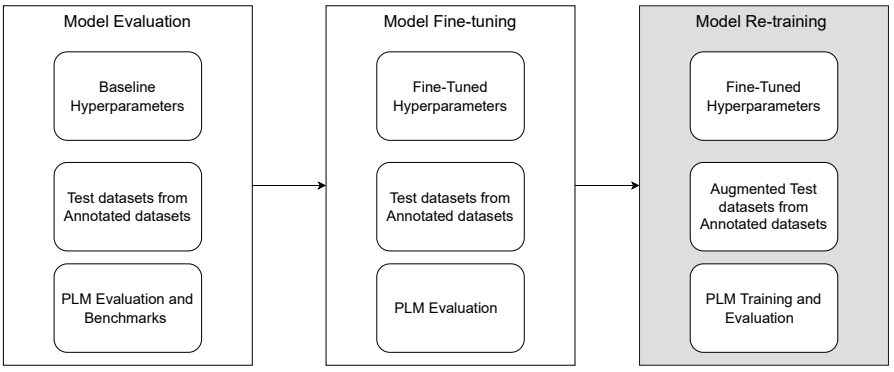

FIGURE 4.1: Pre-trained Model process

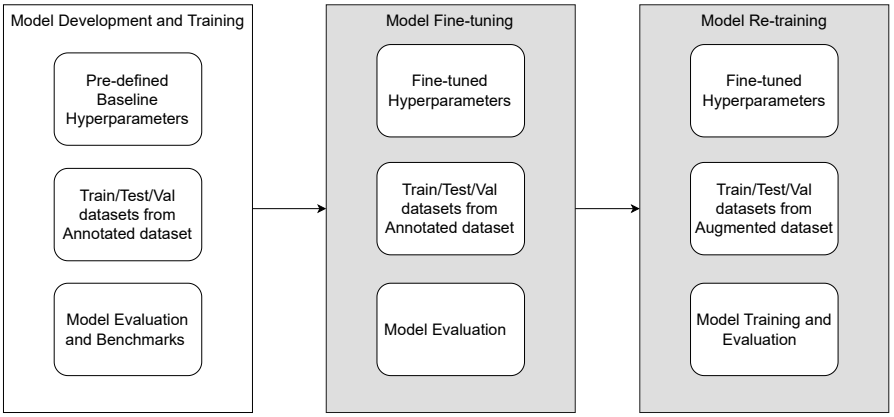

FIGURE 4.2: Siamese Network process

Table 4.3 provides a summary of the training datasets, each corresponding to the respective PLM.

Model	AfriSenti (swah)	DSFSI (zul)	DSFSI (tsn)	Code-mixed
AfriBERTa	x			
AfroXLMR (base)	x	x		x
AfroLM	x	x	x	
PuoBERTa			x	

TABLE 4.3: The different training datasets corresponding to the relevant PLM

4.3.1 AfriBERTa

AfriBERTa is one of the first models to be developed by training it on purely low-resouce languages. At the time of it's development it was believed that training a multilingual model would require joint training with high-resource language, however Ogueji et al. [2021] debunked this theory and proved that multilingual models can be trained using purely low resource languages and furthermore, they can be trained using much smaller datasets that expected, i.e. datasets that are less than 1 GB of text.

AfriBERTa, covering 11 African languages as outlined in Table 4.1, was employed for sentiment prediction in our Swahili dataset. Initially, its benchmark performance was assessed using the AfriSenti Swahili dataset [Muhammad et al., 2023a], yielding an F1-score of 0.461. The ROC curves, depicted in Figure 4.3, reveal a nuanced performance: while the curves for neutral (class 0) and negative (class 1) sentiments are somewhat favorably positioned in the top left corner, indicating promising model accuracy, the curve for the positive (class 2) sentiment lags in the lower right section, signaling a need for improvement in positive sentiment predictions. Given these results, it became evident that model fine-tuning was required to enhance performance. Supporting literature, such as the study by Myoya et al. [2023], confirmed the general efficacy of fine-tuning in boosting PLM performance. However, our approach extended beyond fine-tuning and incorporated additional measures such as data augmentation.

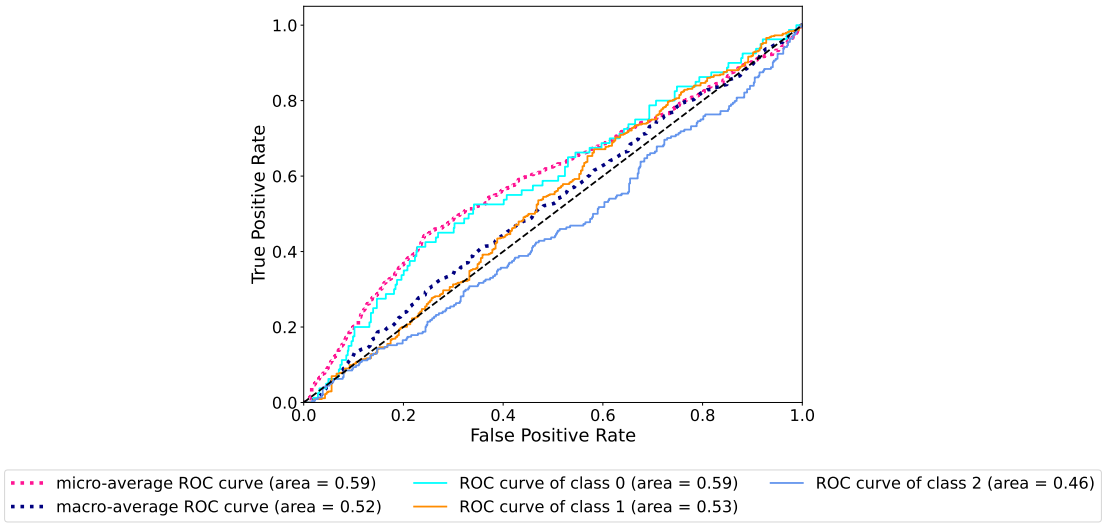

FIGURE 4.3: Benchmark AfriBERTa performance on classic three level sentiment analysis using the AfriSenti Swahili dataset

The fine-tuning process, characterised by iterative experimentation and optimisation methods such as grid search, aimed to identify the most effective hyperparameter combination for maximising the F1-score. The *HiPlot* library [Facebook Research, 2023] was used to visualise the experimentation logs extracted from the Weights&Biases platform [Weights & Biases, Inc., 2023] used in the model training and evaluation process (see Appendix A for visualisations). The resulting hyperparameters, detailed in Table 4.4, led to a substantial 37% improvement in model performance. The fine-tuned ROC curves, as shown in Figure 4.4, also demonstrated enhanced classification efficacy across all sentiment labels.

TABLE 4.4: AfriBERTa Model Hyperparameters applied when evaluating the AfriSenti Swahili dataset

Hyperparameter	Baseline	Fine-tuned
Per Device Train Batch Size	32	16
Per Device Eval Batch Size	32	16
Epochs	2	2
Weight Decay	0	0.02839
Seed	999	16978
Learning Rate	0.0001	5e-5
Adafactor	True	True
Adam β_1	0.9	0.7640
Adam β_2	0.999	0.7439
Adam ϵ	1e-8	3e-8
Max Gradient Norm	1	0.4773
Metric for best model	'eval_loss'	'eval_loss'
Gradient Accumulation Steps	8	1
Warm up steps	40000	0
Dataloader num of workers	6	4
F1-Score	**0.46**	**0.61**

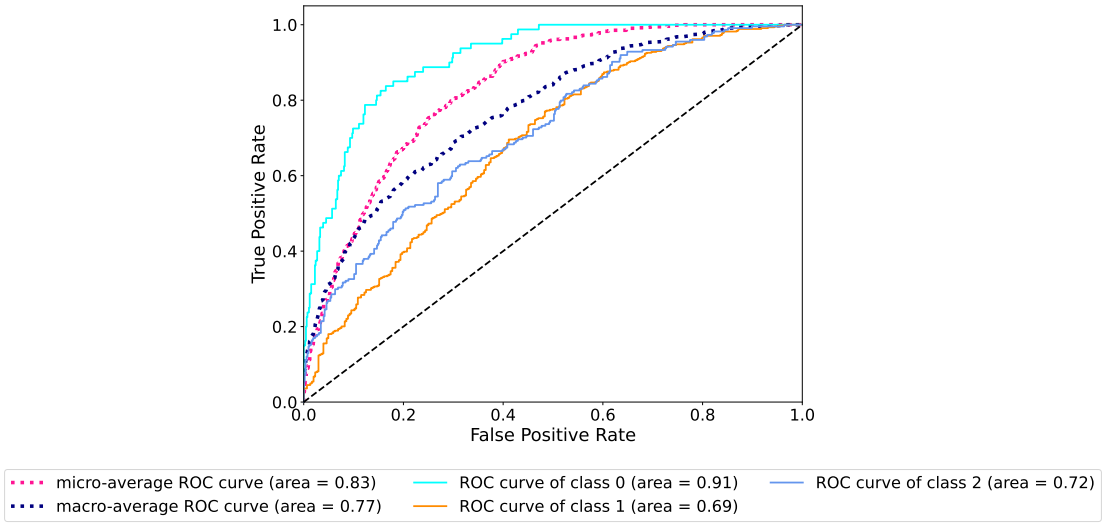

FIGURE 4.4: Fine-tuned AfriBERTa performance on classic three level sentiment analysis

Following the initial evaluation, data augmentation techniques were implemented, and the model underwent re-evaluation using the augmented datasets derived from the AfriSenti (swa) dataset (refer to Table 4.2). The F1-scores obtained after the application of data augmentation are depicted in Figure 4.5.

The outcomes demonstrated a stagnation in model performance, with F1-scores plateauing at 0.63 even after the application of data augmentation. This suggests that the augmentation techniques employed might not have introduced sufficient variability into the dataset. Consequently, enhancing model performance in our context likely required the infusion of additional data to foster greater diversity. This process was anticipated to evolve continuously in our future work through the ongoing validation and retraining of the model, as it processed and learned from new data extracted from social media platforms.

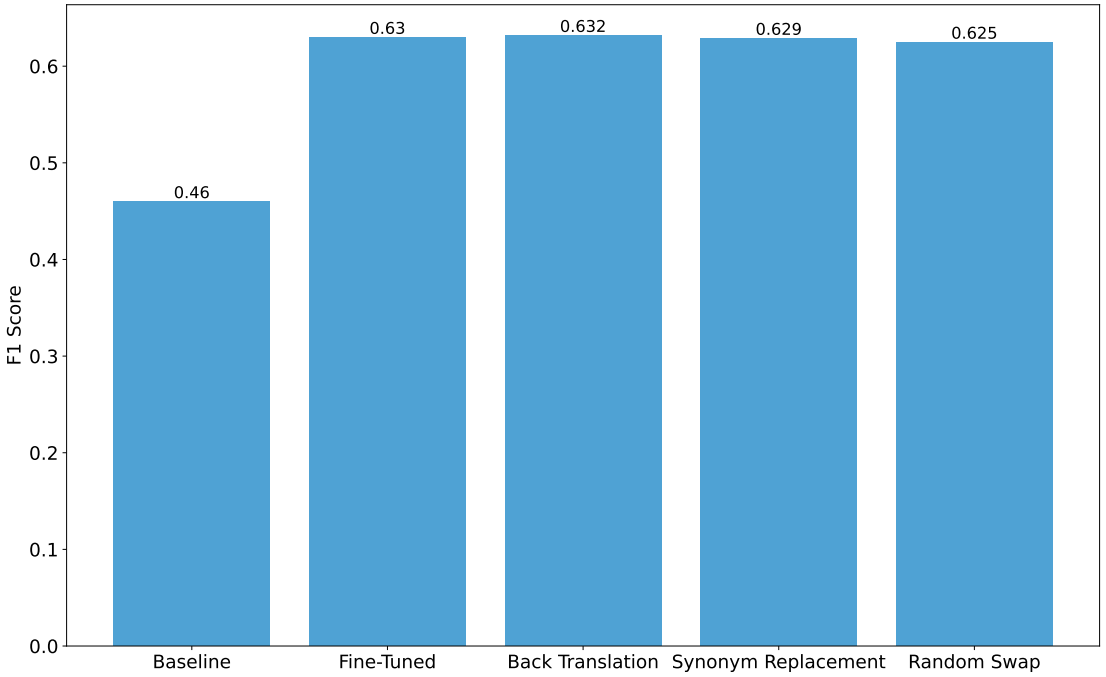

FIGURE 4.5: AfriBERTa Average F1-Score

4.3.2 AfroXLMR (base)

The AfroXLMR model represents a significant advancement in the development of multilingual pre-trained language models (PLMs), particularly for African languages which often face challenges due to being low-resourced and unseen during pre-training. This model utilises the concept of Multilingual Adaptive Fine-Tuning (MAFT) on 17 well-resourced African languages and three other major languages prevalent in Africa [Alabi et al., 2022]. This approach not only fostered cross-lingual transfer learning but also addressed the limitations of Language Adaptive Fine-Tuning (LAFT), such as excessive disk space usage and reduced cross-lingual capabilities due to language-specific specialisation. By strategically removing non-African script tokens from the embedding layer, AfroXLMR achieved a substantial reduction in model size, approximately 50%, without compromising performance. The model's effectiveness was demonstrated through competitive results in three key NLP tasks, namely, Named Entity Recognition (NER), news topic classification, and sentiment classification, when compared to individual LAFT applications, while also significantly conserving disk space. Furthermore, AfroXLMR enhanced zero-shot cross-lingual transfer capabilities, a crucial feature for parameter-efficient fine-tuning methods.

In light of the insights gained from the performance of AfriBERTa, we decided to initiate the fine-tuning process right from the outset for AfroXLMR, employing the grid search optimisation method to identify and utilise hyperparameters in our experimental procedures. The performances of AfroXLMR on Swahili, isiZulu, and the code-mixed datasets (Swahili-English) are detailed in Table 4.5. Additionally, the ROC curves for each dataset are illustrated in Figures 4.6 to 4.8, providing a visual representation of the model's classification capabilities across these diverse linguistic datasets.

TABLE 4.5: AfroXLMR Model Hyperparameters

Hyperparameter	AfriSenti (swah)	DSFSI (zul)	Code-mixed (eng-swah)
Per Device Train Batch Size	16	16	16
Per Device Eval Batch Size	16	16	16
Epochs	10	10	10
Weight Decay	0.2249	0.2249	0.2249
Seed	18790	18790	18790
Learning Rate	4e-5	4e-5	4e-5
Adafactor	True	True	True
Adam β_1	0.2727	0.2727	0.2727
Adam β_2	0.6734	0.6734	0.6734
Adam ϵ	3e-10	3e-10	3e-10
Max Gradient Norm	0.8040	0.8040	0.8040
Metric for best model	'eval_loss'	'eval_loss'	'eval_loss'
Gradient Accumulation Steps	1	1	1
Warm up steps	0	0	0
Dataloader num of workers	4	4	4
F1-Score	**0.59**	**0.82**	**0.97**

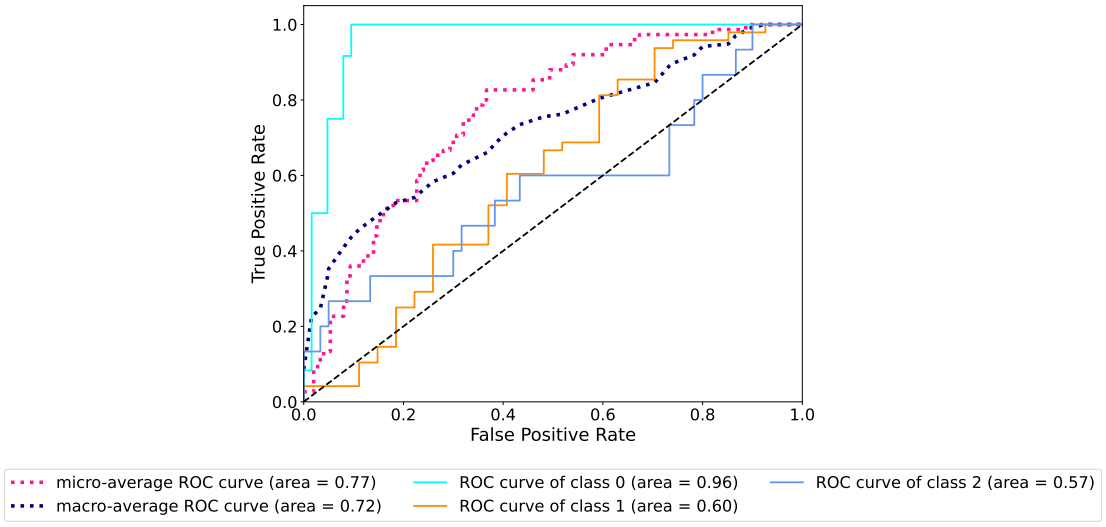

FIGURE 4.6: ROC curve of AfroXLMR: AfriSenti (swah) dataset

The analysis of the results revealed distinct challenges faced by the model with the different datasets. Specifically, with the Swahili dataset, the model exhibited difficulties in accurately identifying neutral labels. In contrast, for the isiZulu dataset, the primary challenge lay in the correct classification of negative labels. The performance on the code-mixed dataset, which incorporated English, a high-resource language, was predictably strong. This outcome aligned with expectations, considering the model's foundation on X-LMR [Conneau et al., 2019], a model pre-trained on English datasets. The proficiency in handling English-based data highlights the benefits of extensive pre-training. These insights informed our exploration of leveraging high-resource languages in model training, particularly as we delved into the development of the Siamese Neural Network in Section 4.3.5.

Considering the findings obtained from AfriBERTa, a discussion emerged about the potential impact of data augmentation on the model's performance. We weighed the benefits of investing additional resources into re-training and evaluating the model post data augmentation, especially

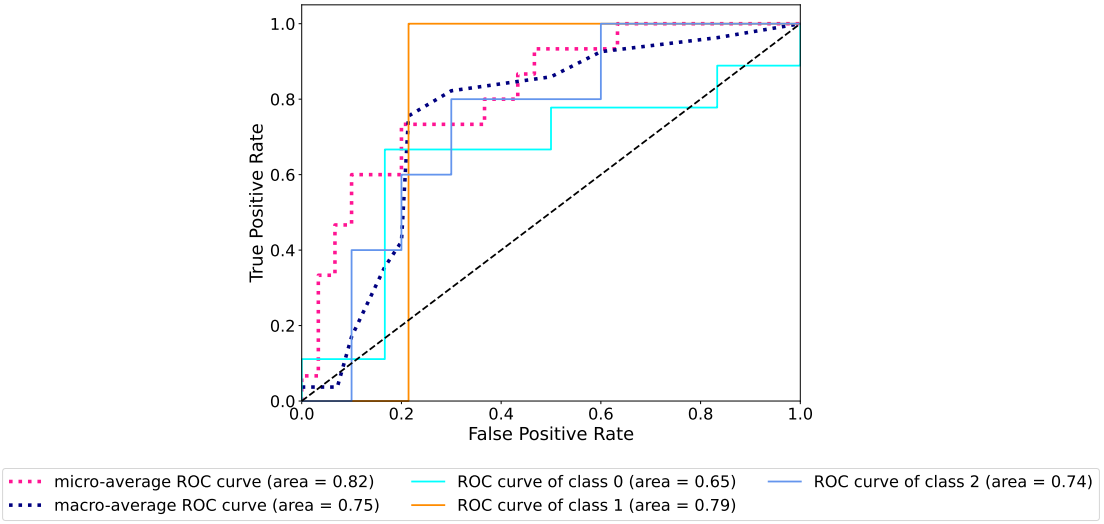

FIGURE 4.7: ROC curve of AfroXLMR: DSFSI (zul) dataset

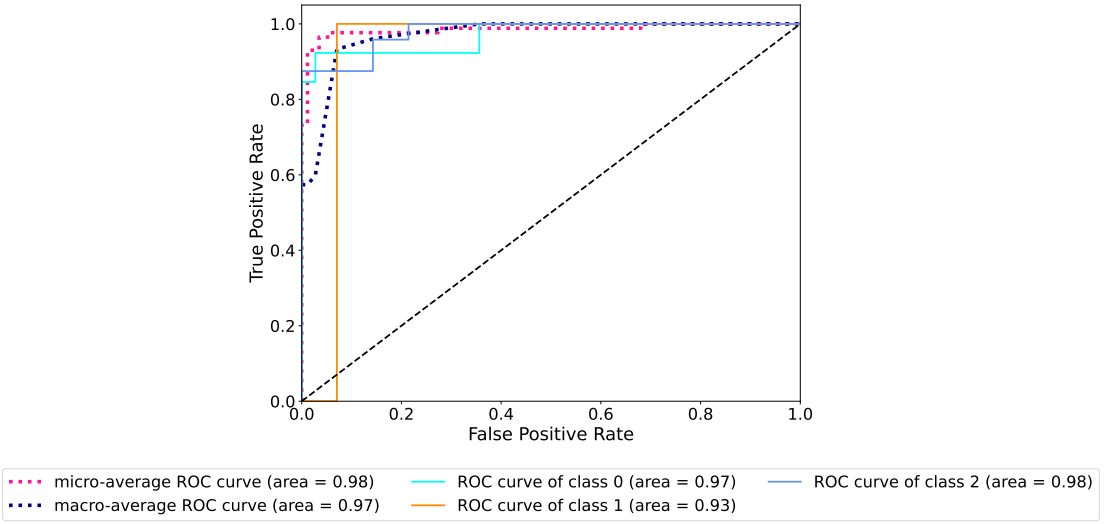

FIGURE 4.8: ROC curve of AfroXLMR: Code-mixed (eng-swah) dataset

in light of the potential for minimal or negligible performance enhancements. This consideration was based on the observation that the data augmentation procedures implemented did not introduce sufficient variability into the dataset. Therefore, the decision to re-train the model using the augmented data hinged on a careful assessment of resource investment against the anticipated improvements in model accuracy and robustness. Ultimately, it was decided not to proceed with retraining and evaluating the model on the augmented data.

4.3.3 AfroLM

AfroLM emerged as a groundbreaking multilingual pre-trained language model, specifically tailored to address the challenges faced by African languages in the field of Natural Language Processing (NLP). These challenges often often being the scarcity of training data, which is a critical resource for developing large multilingual models. AfroLM leveraged the concept of active learning, which is a semi-supervised learning algorithm that dynamically identifies the

most beneficial training samples, thereby optimising performance on downstream NLP tasks. This approach was particularly effective in mitigating real-world data scarcity issues. AfroLM, pre-trained from scratch on 23 African languages, represented the most extensive effort in the domain of active learning in NLP at the time. Utilising a novel self-active learning framework and trained on a dataset significantly smaller (14 times) than existing baselines, AfroLM demonstrated superior performance over several established multilingual models (such as AfriBERTa, XLMR-base, and mBERT) across a variety of NLP tasks including Named Entity Recognition (NER), text classification, and sentiment analysis. Furthermore, its ability to generalise effectively across different domains was evidenced through additional out-of-domain sentiment analysis experiments, showcasing its robustness and versatility.

Following the same methodology used in the evaluations of AfriBERTA and AfroXLMR, as detailed in Sections 4.3.1 and 4.3.2, the hyperparameters and ROC curves of the model are presented in Table 4.6 and Figures 4.9 through 4.11.

TABLE 4.6: AfroLM Model Hyperparameters

Hyperparameter	AfriSenti (swah)	DSFSI (zul)	DSFSI (tsn)
Per Device Train Batch Size	16	16	16
Per Device Eval Batch Size	16	16	16
Epochs	10	10	10
Weight Decay	0.0494	0.0494	0.0494
Seed	14640	14640	14640
Learning Rate	5e-5	5e-5	5e-5
Adafactor	True	True	True
Adam β_1	0.5542	0.5542	0.5542
Adam β_2	0.9000	0.9000	0.9000
Adam ϵ	3e-8	3e-8	3e-8
Max Gradient Norm	0.1338	0.1338	0.1338
Metric for best model	'eval_loss'	'eval_loss'	'eval_loss'
Gradient Accumulation Steps	1	1	1
Warm up steps	0	0	0
Dataloader num of workers	4	4	4
F1-Score	**0.58**	**0.59**	**0.62**

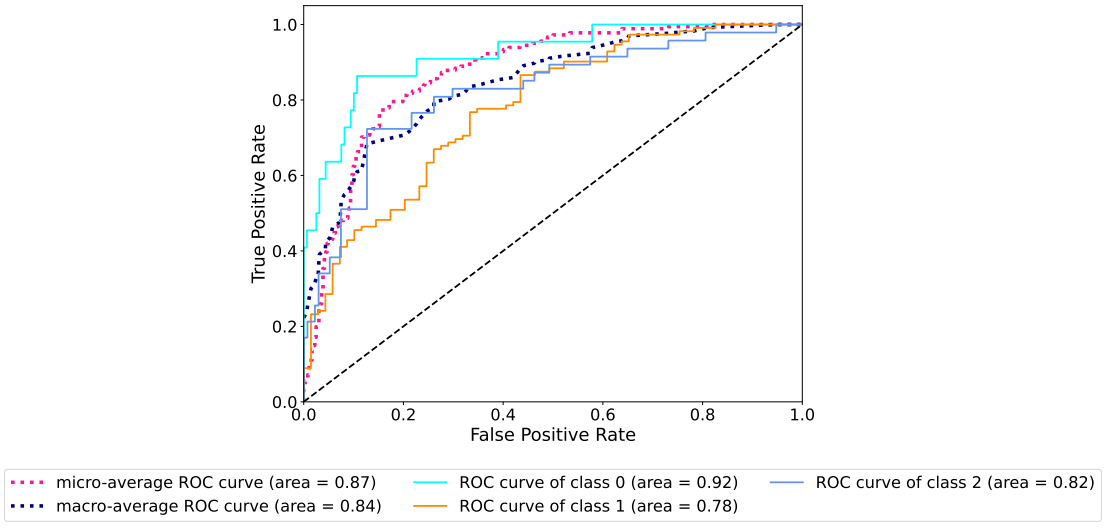

FIGURE 4.9: ROC curve of AfroLM: AfriSenti (swah) dataset

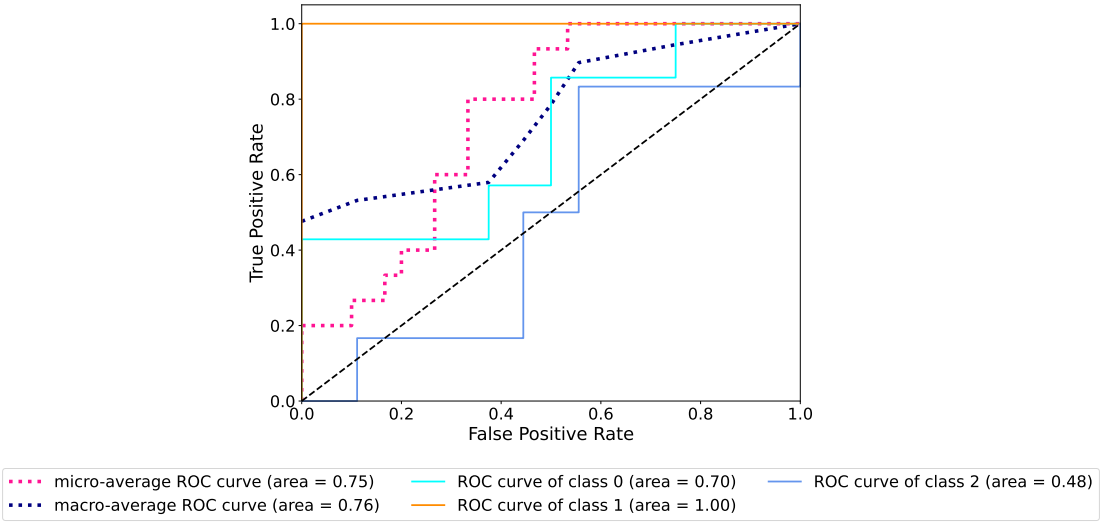

FIGURE 4.10: ROC curve of AfroLM: DSFSI (zul) dataset

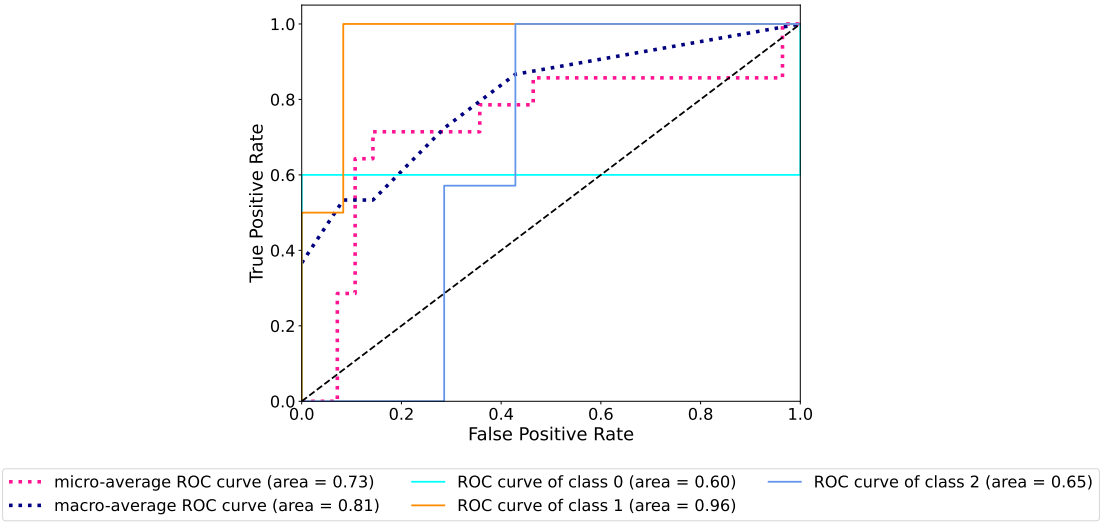

FIGURE 4.11: ROC curve of AfroLM: DSFSI (tsn) dataset

The results demonstrate that the model achieved its best performance with the Swahili dataset, as reflected in both the F1-Score and the ROC curves. The less favorable performance observed with the isiZulu and SeTswana datasets is largely due to constraints in training data availability. Despite this limitation, the model still managed to perform reasonably well, further underscoring its effectiveness in handling data scarcity. Overcoming this challenge would require the incorporation of new data and the subsequent retraining of the model on this newly annotated dataset, a prospect that presents an opportunity for future research and development.

4.3.4 PuoBERTa

The PuoBERTa model is a customised masked language model trained specifically for SeTswana. The model was trained on monolingual sources such as NCHLT Setswana [Eiselen and Puttkammer, 2014] corpus, the South African Constitution 10 [Republic of South Africa, 1996], the Leipzig Setswana BW, ZA corpora [Goldhahn et al., 2012] and more recent corpora such as the Vuk'zenzele Setswana Corpora [Lastrucci et al., 2023] and South African Cabinet Speeches [Adelani et al., 2023]. The model was evaluated on NLP downstream tasks such as part-of-speech (POS) tagging, named entity recognition (NER), and news categorisation. In the context of this study, PuoBERTa was applied to sentiment classification, thereby establishing a benchmark for its performance in this specific domain. This application provides valuable insights into the model's capabilities in interpreting and analysing sentiments, a crucial aspect of language understanding.

TABLE 4.7: PuoBERTa Model Hyperparameters

Hyperparameter	DSFSI (tsn)
Per Device Train Batch Size	16
Per Device Eval Batch Size	16
Epochs	2
Weight Decay	0.02839
Seed	16978
Learning Rate	5e-5
Adafactor	True
Adam β_1	0.7640
Adam β_2	0.7439
Adam ϵ	3e-8
Max Gradient Norm	0.4773
Metric for best model	'eval_loss'
Gradient Accumulation Steps	1
Warm up steps	0
Dataloader num of workers	4
F1-Score	**0.43**

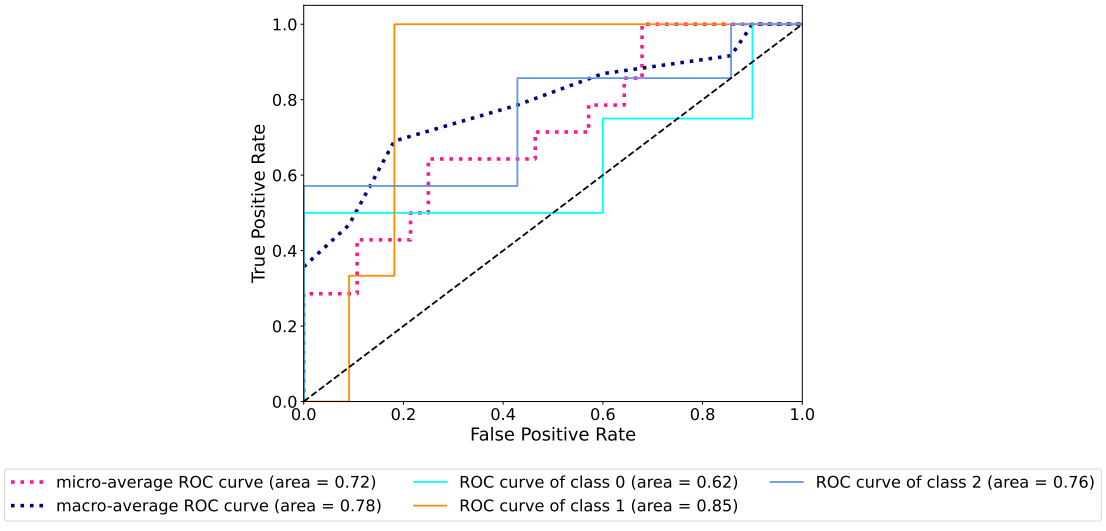

FIGURE 4.12: ROC curve of PuoBERTa: SeTswana dataset

The results are encouraging, demonstrating that the model achieved a respectable level of performance in sentiment classification, especially considering this was its first application in such a downstream task. A detailed examination of the ROC curves revealed that the model encountered difficulties in accurately classifying negative and neutral labels. As this represents the benchmark performance of the model in the realm of sentiment classification, it certainly shows promise. However, to realise its full potential, the model requires further training specifically focused on improving its proficiency in sentiment classification.

4.3.5 Siamese Neural Network

In our study, we adopted the innovative approach of developing a Siamese Neural Network, drawing inspiration from the work of Choudhary et al. [2018]. Their method, Sentiment Analysis of Code-Mixed Text (SACMT), effectively classified sentiments of code-mixed data sourced from social media platforms. The Siamese network operates by pairing two textual inputs, projecting them onto a shared sentiment classification space, and then measuring the similarity between the resulting vectors. The underlying principle is straightforward: text inputs with identical sentiments should exhibit a smaller similarity metric, while those with differing sentiments should show a larger one. This approach was particularly advantageous for sentiment analysis across languages of varying resource levels, as it allowed for leveraging annotated corpora from high-resource languages.

The proposed architecture of our Siamese network, as illustrated in Figure 4.13, incorporated two input tensors capable of handling dynamic sequence lengths and 300 features, typically word embeddings. These inputs then underwent processing through shared dense neural network layers, each consisting of 128 units with ReLU activation. The term 'shared' here implied identical weights and biases for both input pathways. Independently navigating through these shared layers, each input tensor would then produce an output tensor. The network would then then employ a metric like Cosine similarity to compute a similarity score between these outputs, with the network's loss function playing a pivotal role in this training process. This function would be used to penalise pairs that were incorrectly categorised as either too similar or too dissimilar. An Adam optimiser, set at a predefined learning rate, was proposed for loss minimisation.

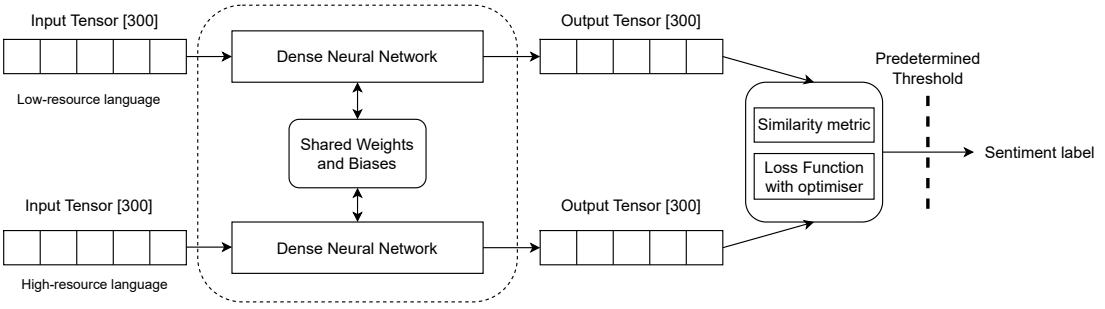

FIGURE 4.13: Siamese Network Architecture

The crux of the Siamese Network was to fine-tune network parameters to minimise the symmetric similarity metric between the output tensors. This entailed employing a loss function during training, enabling the network to learn to map sentences with similar sentiments closer in

sentiment space, while distancing those with disparate sentiments. After constructing the model architecture we utilised our available annotated data, which included the code-mixed dataset paired with the AfriSenti Swahili dataset to create input pairs. The pairings were designed to match high-resource (code-mixed) and low-resource (Swahili) inputs. The pairs were further classified into matching (identical sentiment labels) or non-matching pairs, with the latter labeled as "non-matching" to represent an "unknown" sentiment outcome. Thereafter, the pairs dataset was divided into training, testing, and validation sets, maintaining an 80/10/10 split. The baseline results from the model training and evaluation are depicted in Figure 4.14, and the model hyperparameters are detailed in Table 4.8.

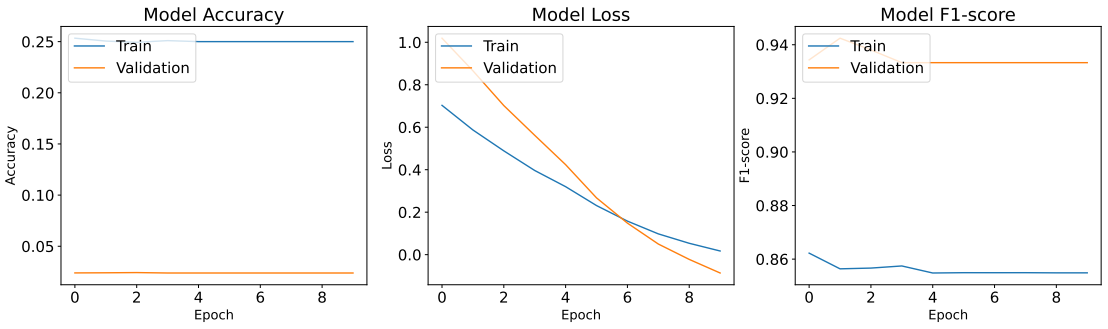

FIGURE 4.14: Siamese Network Training and Evaluation Results.

TABLE 4.8: Siamese Network Hyperparameters

Hyperparameter	Input pairs (swah-eng & swah)
Loss Function	Energy Function
Similarity Score	Cosine Similarity
Epochs	10
Learning Rate	5e-5
Batch Size	32
Accuracy	**0.25**
F1-Score	**0.93**

The results indicated a high F1-Score but low accuracy. In light of the discussion in the article by [Shmueli, 2019], this discrepancy may be attributed to class imbalance or potentially an overly complex model architecture for the data. Additionally, the confusion matrix in Figure 4.15 reveals the model's bias towards predicting negative sentiment pairs, suggesting either an over-representation of negative labels or, again, an issue with the model's complexity. Given that SMOTE was conducted to cater for the class imbalance issue, we assumed model complexity to be cause of the poor performance.

Faced with these challenges, we decided to simplify the model to focus on distinguishing between matching and non-matching sentiment pairs. Consequently, the Siamese architecture was reconfigured to assess the similarity between two input samples, determining the sentiment label based on the distance to a reference input (a labeled, high-resource input) from the new input (potentially a low-resource language). We established a distance threshold of 0.5 for this classification process. The modified network showed improved results in terms of F1-score (0.67) and accuracy (0.53), but further fine-tuning is necessary to achieve optimal performance. Future work will focus on continuing this development and fine-tuning process.

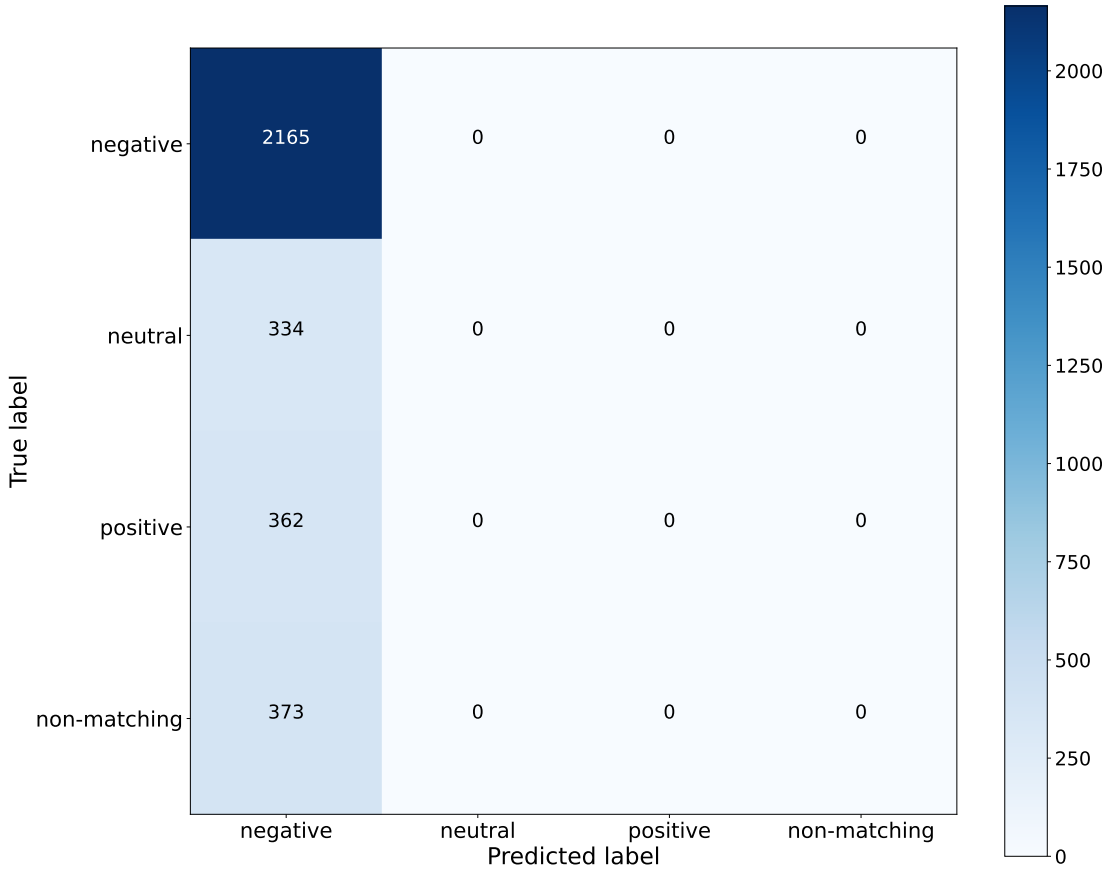

FIGURE 4.15: Siamese Network Confusion matrix.

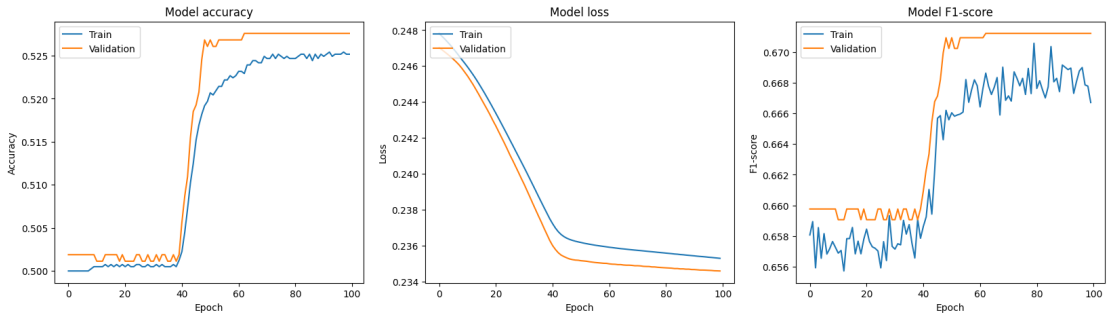

FIGURE 4.16: Simplified Siamese Network Training and Evaluation results.

It should be noted that while the Siamese network approach is powerful for discerning data similarities, for direct sentiment classification of individual tweets, more straightforward neural network architectures like standard LSTM or CNN, trained on labeled sentiment data, might be more effective and easier to implement. However, the Siamese method's advantage, as highlighted by Choudhary et al. [2018], lies in its ability to exploit relationships between different data points (e.g., code-mixed and pure English tweets) to deepen sentiment understanding.

The ultimate goal during training was to ensure that tweets with similar sentiments were proximal in sentiment space, while those with differing sentiments were distanced. With adequate pairs and iterations, the Siamese network is expected to become more adept at discerning between similar and different sentiments in tweets. In practice, after training, the introduction of a new

pair of tweets to the trained network would yield a distance value, indicating whether the tweets share similar or dissimilar sentiments, based on their proximity in sentiment space.

4.4 Comparison of commuter sentiment and public transport provider ratings

This section of our study was dedicated to assessing the effectiveness of our models in a real-world application: predicting commuter sentiment towards public transport in three selected countries, namely Kenya, Tanzania, and South Africa. For this task, we chose the models that not only demonstrated the highest F1 scores but also maintained high accuracy. These were AfriBERTa for Swahili, AfroXLMR-base for isiZulu and code-mixed data, and AfroLM for SeTswana. The model F1-score are summarised in Table 4.9.

Model	Swahili (swh)	isiZulu (zul)	Setswana (tsn)	Code-mixed (with eng)
AfriBERTa	0.61	-	-	-
AfroXLMR (base)	0.56	0.83	-	0.97
AfroLM	0.58	0.59	0.62	-
PuoBERTa	-	-	0.43	-

TABLE 4.9: Model Evaluation (F1-Score)

In South Africa, the Railway Safety Regulator (RSR) annually reports on railway safety, adhering to the South African National Standards (SANS) categories SABS [2009]. These reports encompass statistics on security incidents, operational occurrences, and incidents involving trains striking individuals Regulator [2021]. However, with the decline of the PRASA Metrorail service, there's a growing concern about under-reporting, making the safety data less reliable. To bridge this information gap between service providers and commuters, supplementary data sources like opinion mining can be invaluable.

When considering the sentiment of commuters based on tweets related to rail services in South Africa, the predominant sentiment was negative, as illustrated in Figure 4.17. This sentiment corresponded with themes derived from the feature extraction process that related to destruction, failure and vandalism. This sentiment corresponds to the rating of the RSR where the ratio of security-related incidents to operational occurrences has increased steadily since 2017 (corresponding to the same timeline as that of the tweets extracted). See Figure 4.18

Additionally, sentiment analysis was conducted on the subsets of data corresponding to the themes identified in Section 3.2.4. These findings are illustrated in Figure 4.19, revealing a predominantly negative sentiment across the theme subsets. Notably, the theme 'manhlamaza' exhibits a mixture of neutral sentiments, though the tweets under this theme were ambiguous within the context of public transport. This negative trend aligns with expectations, as commuters tend to use Twitter primarily to voice complaints about services rather than to share positive experiences, as outlined in the study by Qi et al. [2020].

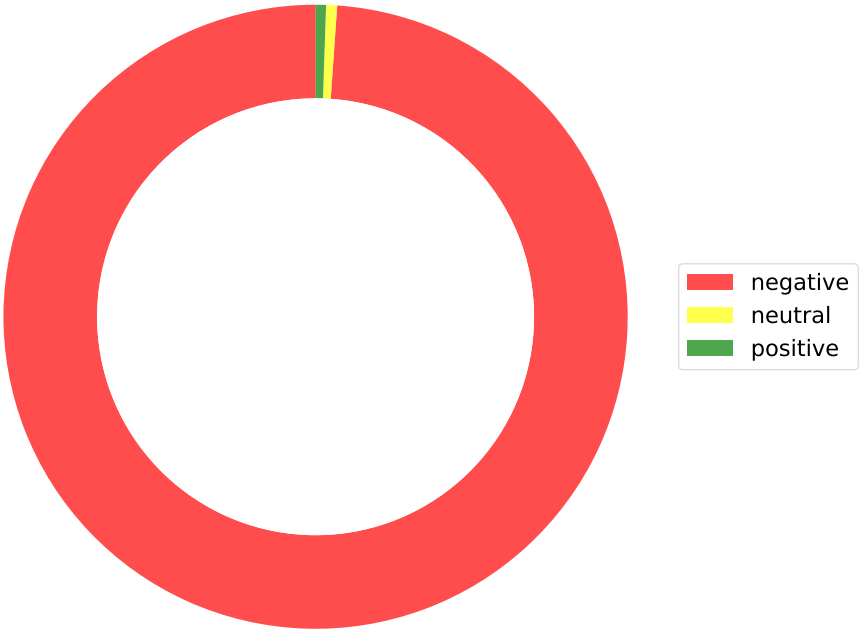

FIGURE 4.17: Sentiment distribution of South African tweets related to train usage

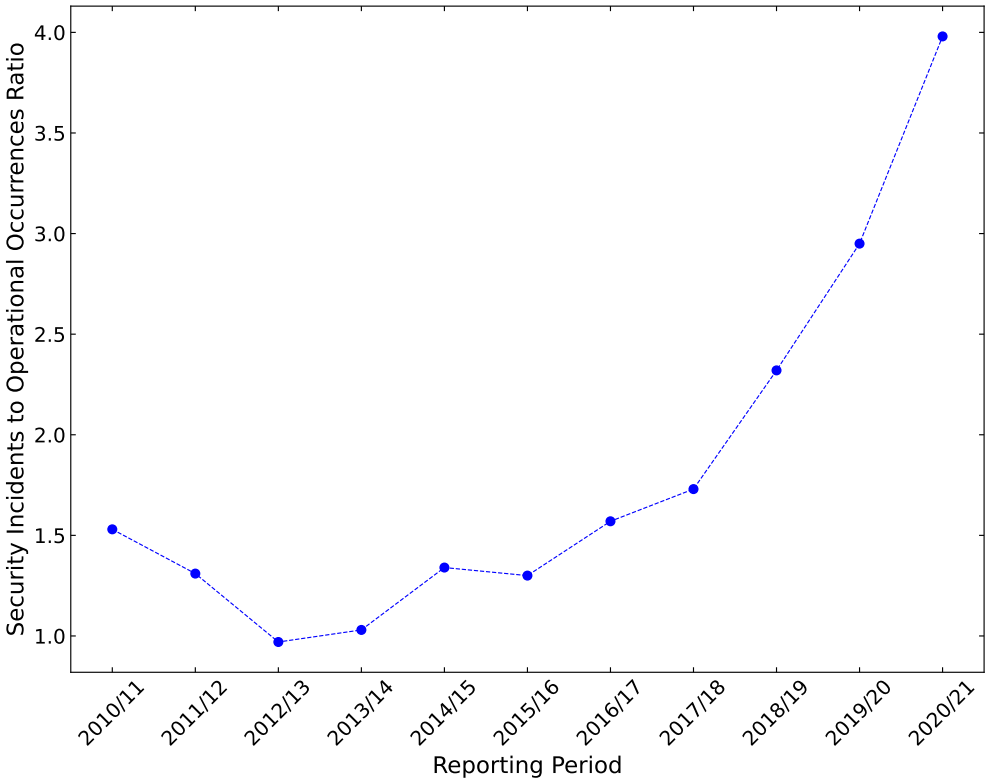

FIGURE 4.18: RSR ratio of security-related incidents to operational occurrences [Regulator, 2021]

In Kenya, the *Matatu* industry, which represents the dominant minibus taxi sector, has been the focus of reform efforts since 2004 Githui et al. [2009]. These reforms were designed to

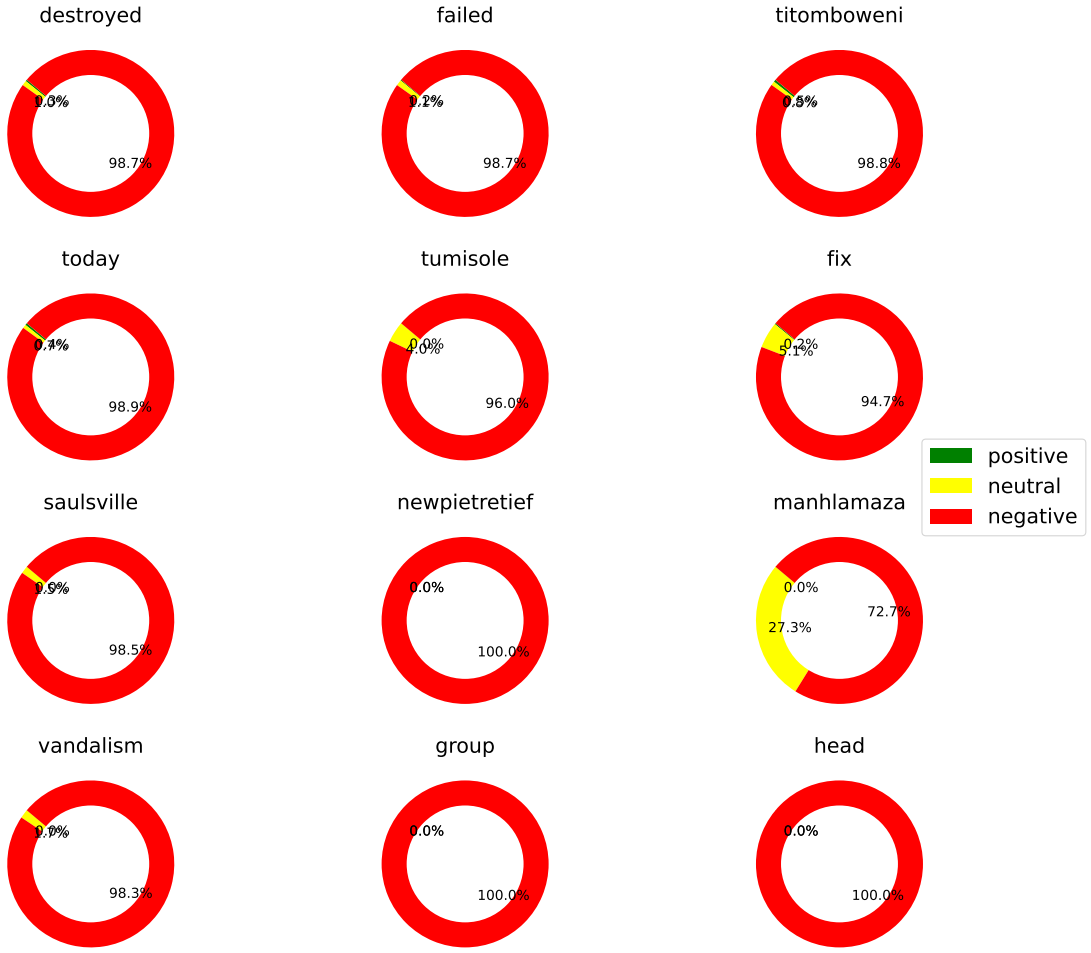

FIGURE 4.19: Sentiment distribution of South African tweets according to the themes derived from Section 3.2.4

mitigate over-speeding accidents, bolster commuter safety, ensure the competence of drivers and conductors, eliminate unauthorised personnel, and regulate vehicle operations Chitere and Kibua [2004]. Despite stringent enforcement of the Traffic Act 403, *Matatus* continue to be a leading cause of road accidents Mburu [2023]. Additionally, the industry has been marred by incidents of commuter harassment, especially targeting women Mwaura [2020]. Leveraging opinion mining for incident detection can offer near real-time insights and actionable solutions. An analysis of tweets concerning *Matatus* revealed a predominantly negative sentiment, as illustrated in Figure 4.20. The primary themes extracted included unpredictable price hikes, over-speeding, violent crime, and safety concerns. The World Health Organisation reported that, for individuals aged between five and 70 in Kenya, road-related fatalities rank among the top five causes of death WHO [2022]. Given that *Matatus* play a significant role in these accidents, reforms in the industry could substantially reduce the loss of productive citizens.

Similarly to the analysis of South African tweets, sentiment analysis was performed on data subsets related to the themes identified in Section 3.2.4. The predominant sentiment, as depicted in the results showcased in Figure 4.21, was negative, echoing the patterns observed in Figure 4.20. Nevertheless, instances of positive sentiments were identified within the themes of 'heard', 'ina', 'ply', 'madawa', and 'somewhere'. These themes encompass discussions related to Kenya

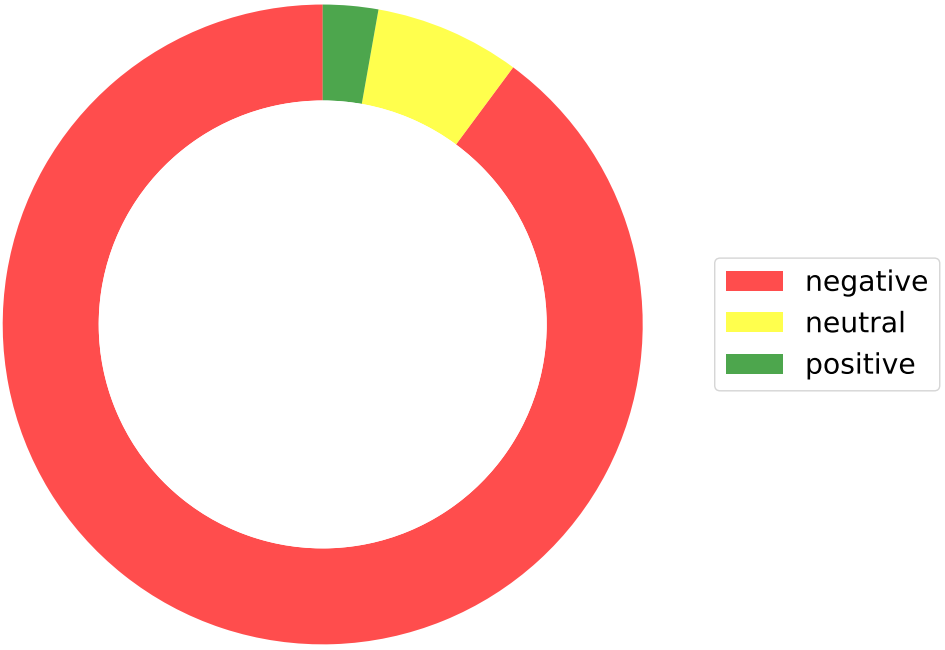

FIGURE 4.20: Sentiment distribution of Kenyan tweets related to *Matatu* usage

Power, pharmaceuticals, and narratives recounting various incidents, including those involving government actions and social gatherings.

FIGURE 4.21: Sentiment distribution of Kenyan tweets according to the themes derived from Section 3.2.4

In Tanzania's Dar es Salaam, the recent launch of the Bus Rapid Transit (BRT) system signified a major effort to enhance the Quality of Service (QoS) in public transport. However, questions emerged regarding its impact on commuter experience and its alignment with user demands. A study by Joseph et al. [2021] found that while the BRT alleviated traffic congestion, issues like prolonged waiting times, overcrowding, and safety concerns persisted. An analysis of related tweets indicated a generally positive sentiment towards bus transport (see Figure 4.22). However, a closer look at the data revealed that many of these tweets were promotional, using positive descriptors such as "Luxurious" and "affordable". Thus, a more authentic source of commuter feedback for Tanzanian travelers would be necessary.

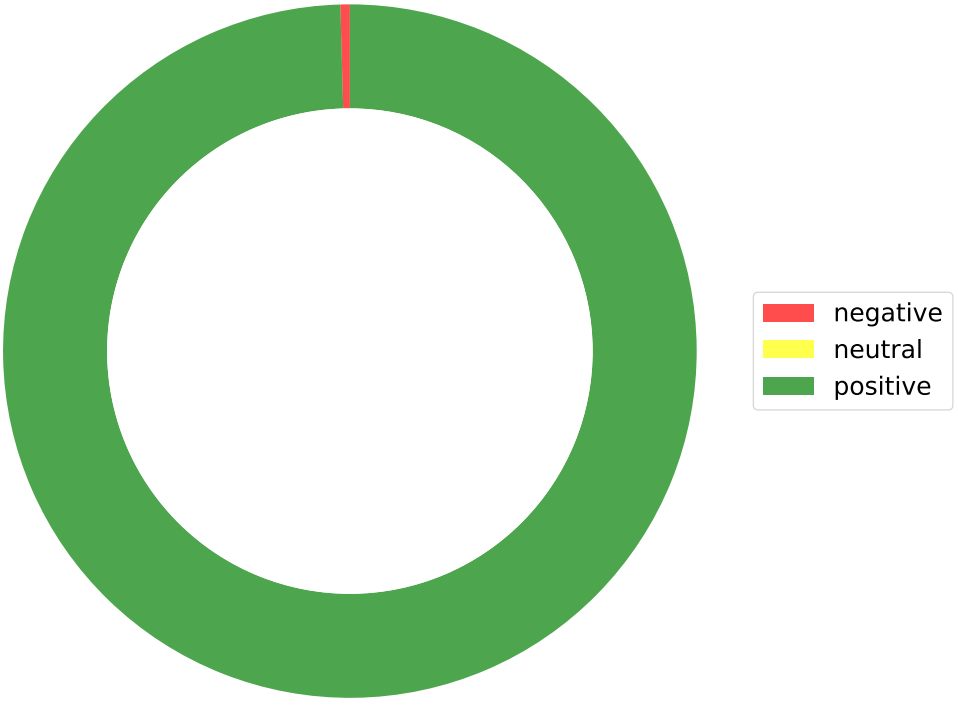

FIGURE 4.22: Sentiment distribution of Tanzanian tweets related to BRT and bus usage

The analysis of the Tanzanian dataset revealed a predominantly positive sentiment across the thematic subsets, as illustrated in Figure 4.23. This finding is consistent with the sentiment trends observed in Figure 4.22, except for the theme 'habari'. This particular theme, which primarily includes tweets related to news reporting, accounted for the presence of a few negative tweets. The overall positive bias identified within these themes can be attributed to the nature of the tweets in the Tanzanian dataset, which were largely focused on advertising.

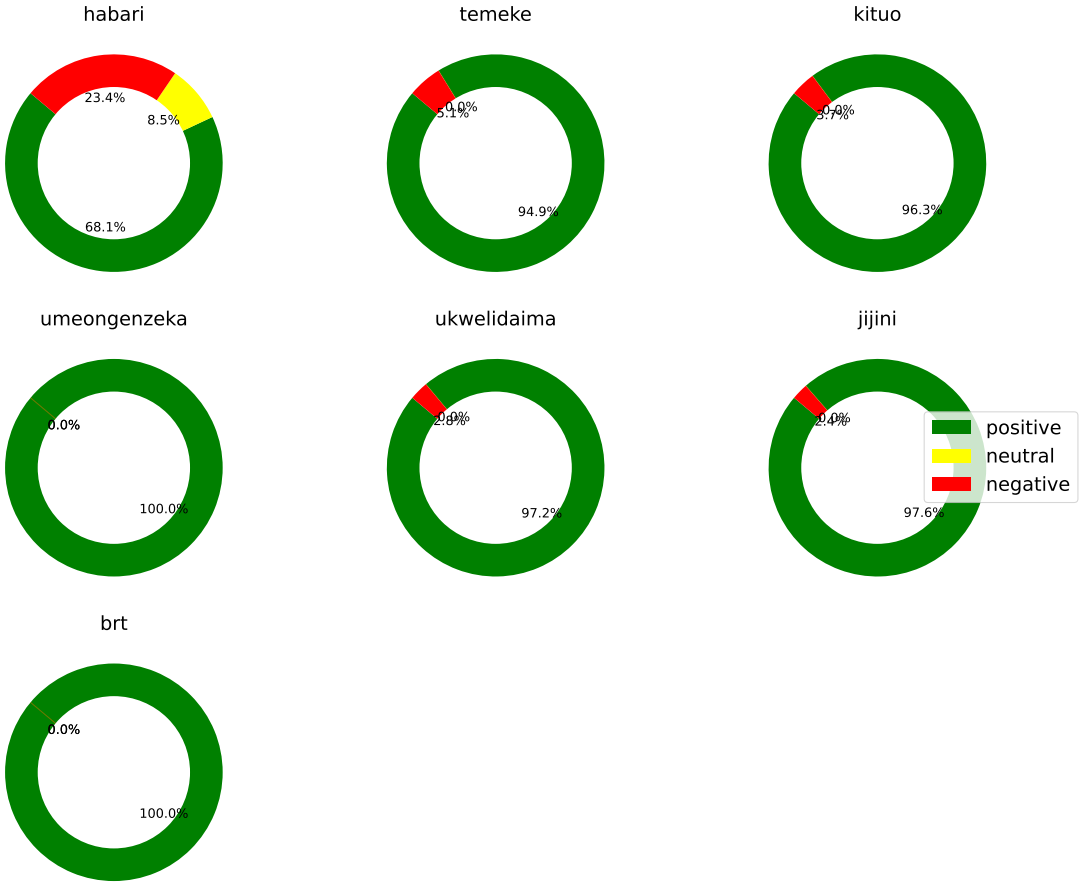

FIGURE 4.23: Sentiment distribution of Tanzanian tweets according to the themes derived from Section 3.2.4

4.4.1 Results Validation

In this chapter, we explored the practical application of Language Models in answering the research question on deriving commuter sentiment from data collected from social media. We explored the implementation of PLMs by testing a few of them to determine which would be best applied in the application of commuter sentiment classification. This application step was carried out in Section 4.4, where the models were chosen on the merit of their performance during their evaluation. However, the question then arises on the reliability of these models whose results would be used in decision-making within the public transport field. Can these model outputs be trusted. To answer this question, manual validation was conducted on 10% of the data classified by each model extracted from each country and the performance was evaluated based on the F1-Score.

TABLE 4.10: Model results validation

	AfriBERTa	AfroXLMR	AfroLM
F1-Score	55.9%	55.5%	60.0%

The results presented in Table 4.10 indicate that the models produce outputs with limited reliability. A closer examination of the sentiment predictions from all three models consistently

showed that the neutral sentiment was often misclassified as negative, as illustrated in Figures 4.24 to 4.26. To rectify this, we plan to continuously sample and annotate data, followed by model retraining. Additionally, we will fine-tune the models by experimenting with their hyperparameters and employ data augmentation techniques to enhance the diversity of the training data.

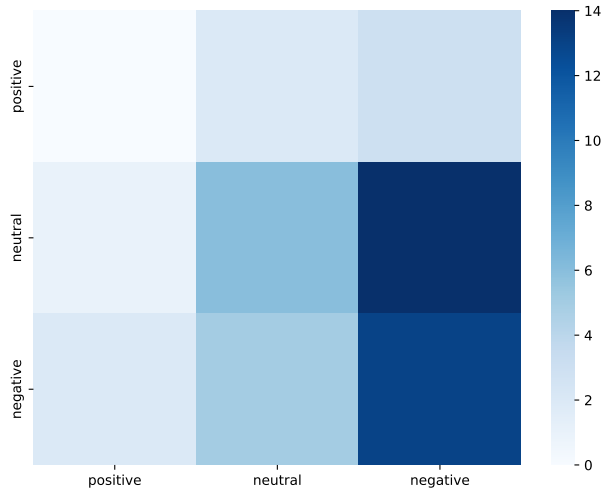

FIGURE 4.24: AfriBERTa Confusion Matrix

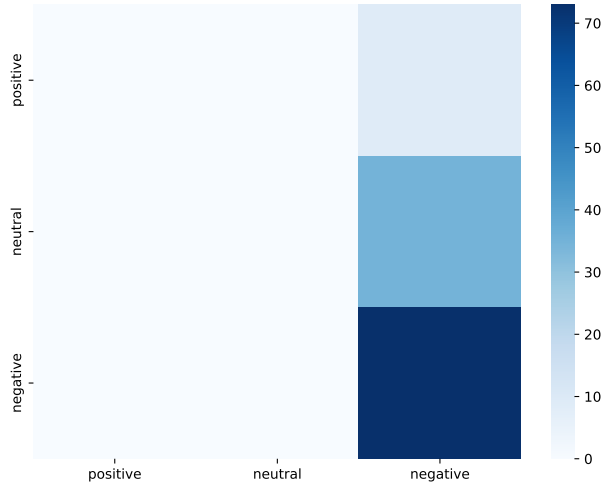

FIGURE 4.25: AfroXLMR Confusion Matrix

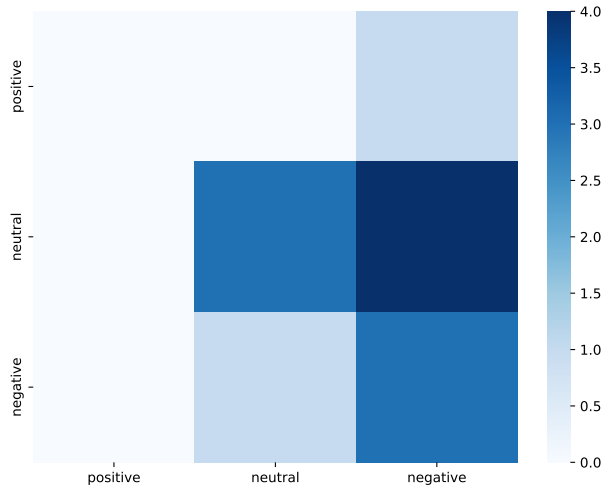

FIGURE 4.26: AfroLM Confusion Matrix

4.5 Discussion

This study presented a comprehensive exploration into the application of various Pre-trained Language Models (PLMs) and the innovative development of a Siamese Neural Network to analyse public transport user sentiment from Twitter data. The study's goal was to understand the user experience across major transport modes in Sub-Saharan countries and identify service gaps and opportunities for improvement.

The PLMs employed included AfriBERTa, AfroXLMR, AfroLM, and PuoBERTa, each tailored to specific African languages or language families. These models underwent extensive evaluation using annotated datasets to establish their efficacy in sentiment prediction. The evaluation revealed varied performance across different languages and datasets, reflecting the unique challenges associated with each language and the availability of training data. AfriBERTa demonstrated promising results, particularly after fine-tuning, whereas AfroXLMR, AfroLM, and PuoBERTa each had their strengths and weaknesses, contingent on the language and dataset applied.

One of the most significant challenges identified was the class imbalance in the datasets, which necessitated the implementation of techniques like SMOTE to ensure a balanced training process. The Siamese Neural Network, developed to address the code-mixed nature of the dataset, provided a novel approach to sentiment analysis. However, its initial complexity led to suboptimal performance, highlighting the need for further fine-tuning and simplification.

Upon simplification, the Siamese Network showed improved results, but further development is required to achieve optimal performance. This finding underscores the need for ongoing research and development in machine learning models, particularly in the context of sentiment analysis for low-resource languages and code-mixed data.

The comparative analysis between the sentiment predictions from these models and the ratings provided by public transport providers offered valuable insights. It revealed discrepancies between the perceived service quality of providers and the actual experiences of users, emphasising the need for more user-centric approaches in public transport services.

In conclusion, this research has demonstrated the potential of NLP and machine learning in providing critical insights into public transport user sentiment. It highlighted the challenges and opportunities in sentiment analysis for African languages and the importance of developing robust, scalable, and adaptable models to cater to the diverse linguistic landscape of Sub-Saharan countries. The study's findings pave the way for future research in this field, with a focus on enhancing model performance, addressing data scarcity, and improving the understanding of user experiences in public transport systems.

Chapter 5

Conclusions

5.1 Conclusions

The following conclusions were drawn from the study:

- **Innovative Application of NLP in Public Transport Sentiment Analysis:** The study demonstrates the successful application of Natural Language Processing (NLP) techniques to analyse public transport user sentiment. This approach highlights the potential of leveraging social media data to gain insights into user experiences and service gaps in public transport systems, particularly in Sub-Saharan Africa.
- **Multilingual Sentiment Analysis Challenges and Solutions:** The research underlines the unique challenges of multilingual sentiment analysis, especially with under-resourced languages like Swahili, Setswana, and isiZulu. The study's use of pre-trained language models (PLMs) like AfriBERTa, AfroXLMR, AfroLM, and PuoBERTa showcases effective strategies to overcome these challenges.
- **Significance of Code-Mixed Data Analysis:** The creation and analysis of a code-mixed dataset, blending English with African languages, reflects the real-world language use in social media. This approach emphasises the importance of including code-mixed data for comprehensive sentiment analysis in multilingual societies.
- **Role of Data Augmentation Techniques:** The application of Synthetic Minority Oversampling Technique (SMOTE) to balance class distribution in training datasets is a pivotal step. It addresses the challenge of class imbalance, a common issue in machine learning datasets, ensuring more reliable and unbiased model performance.
- **Comparative Analysis for Service Improvement:** The study's comparison of sentiment analysis results with public transport provider ratings reveals significant insights. It not only highlights the service quality perceptions but also identifies potential areas for service improvement.
- **Challenges in Automated Sentiment Analysis:** The research indicates limitations in relying solely on automated sentiment analysis, such as auto-labeling based on emojis. It

points out the need for manual validation to ensure the accuracy and reliability of sentiment categorisation.

5.2 Recommendations

The following recommendations are proposed:

- **Enhancing Language Model Training:** Future studies should consider expanding the training datasets for under-resourced languages to enhance the effectiveness of PLMs in sentiment analysis.
- **Diversifying Data Sources:** Incorporating data from a variety of social media platforms can provide a more holistic understanding of public transport user sentiment.
- **Refining Data Augmentation Methods:** While SMOTE is effective, exploring other data augmentation techniques might provide a more nuanced approach to address class imbalance, especially for complex datasets.
- **Continued Manual Validation Efforts:** Due to the nuanced nature of sentiment in social media texts, manual validation remains crucial. Future research should allocate resources for extensive manual validation to complement automated techniques.
- **Applying Findings to Policy and Practice:** The insights from this research should be utilised by public transport authorities and policymakers to inform decisions and improve the quality of service.
- **Exploring Predictive Analytics:** Future research can explore the use of sentiment analysis for predictive analytics in public transport, potentially forecasting user satisfaction and system demands based on sentiment trends.

By addressing these areas, future research can continue to advance the application of NLP in public transport and other domains, particularly in multilingual contexts.

Bibliography

[Adebara et al., 2022] Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*, 2022.

[Adelani et al., 2023] David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. Masakhanews: News topic classification for african languages. *arXiv preprint arXiv:2304.09972*, 2023.

[Alabi et al., 2022] Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, 2022.

[Alexander, 2012] Matthew Alexander. Scotrail and adopt a station: the indirect benefits of community involvement in public transport spaces. *Scottish Transport Review*, 54(April):18–20, 2012.

[Batorsky, 2022] Batorsky. Old problems: The missing voices in natural language processing 2022, Apr 2022. URL https://thegradient.pub/nlp-new-old/.

[Bubeck et al., 2014] Steffen Bubeck, Jan Tomaschek, and Ulrich Fahl. Potential for mitigating greenhouse gases through expanding public transport services: A case study for gauteng province, south africa. *Transportation Research Part D: Transport and Environment*, 32:57–69, 2014.

[Camacho et al., 2016] Tiago Camacho, Marcus Foth, Andry Rakotonirainy, Markus Rittenbruch, and Jonathan Bunker. The role of passenger-centric innovation in the future of public transport. *Public Transport*, 8(3):453–475, 2016.

[Casas and Delmelle, 2017] Irene Casas and Elizabeth C Delmelle. Tweeting about public transit—gleaning public perceptions from a social media microblog. *Case Studies on Transport Policy*, 5(4):634–642, 2017.

[Chawla et al., 2002] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[Chitere and Kibua, 2004] Preston O Chitere and Thomas N Kibua. Efforts to improve road safety in kenya. 2004.

[Choudhary et al., 2018] Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. Sentiment analysis of code-mixed languages leveraging resource rich languages. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 104–114. Springer, 2018.

[Clark and Crous, 2002] Peter Clark and Wilfred Crous. Public transport in metropolitan cape town: Past, present and future. *Transport reviews*, 22(1):77–101, 2002.

[Cndro, 2021] Cndro. How to extract data from facebook using graph api, Sep 2021. URL https://medium.com/@cndro/how-to-extract-data-from-facebook-using-graph-api-ebf4488dee76.

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

[Dossou et al., 2022] Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages, 2022. URL https://arxiv.org/abs/2211.03263.

[Eiselen and Puttkammer, 2014] Roald Eiselen and Martin J Puttkammer. Developing text resources for ten south african languages. In *LREC*, pages 3698–3703. Citeseer, 2014.

[Ekinci et al., 2018] Yeliz Ekinci, Nimet Uray, Füsun Ülengin, and Cem Duran. A segmentation based analysis for measuring customer satisfaction in maritime transportation. *Transport*, 33(1):104–118, 2018.

[Facebook Research, 2023] Facebook Research. Hiplot. https://facebookresearch.github.io/hiplot/, 2023. Accessed: 2023-01-09.

[Githui et al., 2009] John Ngatia Githui, Toshiyuki Okamura, and Fumihiko Nakamura. The structure of users' satisfaction on urban public transport service in developing country: the case of nairobi. In *Proceedings of the Eastern Asia Society for Transportation Studies Vol. 7 (The 8th International Conference of Eastern Asia Society for Transportation Studies, 2009)*, pages 232–232. Eastern Asia Society for Transportation Studies, 2009.

[Goldhahn et al., 2012] Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43, 2012.

[Gupta et al., 2016] Deepak Gupta, Ankit Lamba, Asif Ekbal, and Pushpak Bhattacharyya. Opinion mining in a code-mixed environment: A case study with government portals. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 249–258, 2016.

[Haghighi et al., 2018] N Nima Haghighi, Xiaoyue Cathy Liu, Ran Wei, Wenwen Li, and Hu Shao. Using twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. *Public Transport*, 10:363–377, 2018.

[Helmreich et al., 2004] Stephen Helmreich, David Farwell, Bonnie Dorr, Nizar Habash, Lori Levin, Teruko Mitamura, Florence Reeder, Keith J Miller, Eduard Hovy, Owen Rambow, et al. Interlingual annotation of multilingual text corpora. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 55–62, 2004.

[Joseph et al., 2021] Lucy Joseph, An Neven, Karel Martens, Opportuna Kweka, Geert Wets, and Davy Janssens. Exploring changes in mobility experiences and perceptions after implementation of the bus rapid transit system in dar es salaam. *Case Studies on Transport Policy*, 9(2): 930–938, 2021.

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

[Kanyama, 2004] Ahmad Kanyama. *Public transport in Dar es Salaam, Tanzania: Institutional challenges and opportunities for a sustainable transportation system.* Totalförsvarets forskningsinstitut, Institutionen för miljöstrategiska studier, 2004.

[Keseru et al., 2020] Imre Keseru, Ewoud Heyndels, Tan Dat Ton, and Cathy Macharis. Multitasking on the go: An observation study on local public transport in brussels. *Travel Behaviour and Society*, 18:106–116, 2020.

[Kim et al., 2016] Yoonsang Kim, Jidong Huang, and Sherry Emery. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of medical Internet research*, 18 (2):e41, 2016.

[Kohne et al., 2022] Julian Kohne, Jon D Elhai, and Christian Montag. A practical guide to whatsapp data in social science research. In *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, pages 171–205. Springer, 2022.

[Krizek and El-Geneidy, 2007] Kevin J Krizek and Ahmed El-Geneidy. Segmenting preferences and habits of transit users and non-users. *Journal of public transportation*, 10(3):71–94, 2007.

[Krüger et al., 2021] Fred Krüger, Alexandra Titz, Raphael Arndt, Franziska Groß, Franziska Mehrbach, Vanessa Pajung, Lorenz Suda, Martina Wadenstorfer, and Laura Wimmer. The bus rapid transit (brt) in dar es salaam: A pilot study on critical infrastructure, sustainable urban development and livelihoods. *Sustainability*, 13(3):1058, 2021.

[Kuflik et al., 2017] Tsvi Kuflik, Einat Minkov, Silvio Nocera, Susan Grant-Muller, Ayelet Gal-Tzur, and Itay Shoor. Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies*, 77:275–291, 2017.

[Kuratov and Arkhipov, 2019] Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*, 2019.

[Lastrucci et al., 2023] Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. *arXiv preprint arXiv:2303.03750*, 2023.

[Ledwaba and Marivate, 2022] Mashadi Ledwaba and Vukosi Marivate. Semi-supervised learning approaches for predicting south african political sentiment for local government elections. In *DG.O 2022: The 23rd Annual International Conference on Digital Government Research*, pages 129–137, 2022.

[Mabokela and Schlippe, 2022] Ronny Mabokela and Tim Schlippe. A sentiment corpus for south african under-resourced languages in a multilingual context. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 70–77, 2022.

[Mambina et al., 2022] Iddi S Mambina, Jema D Ndibwile, and Kisangiri F Michael. Classifying swahili smishing attacks for mobile money users: A machine-learning approach. *IEEE Access*, 10:83061–83074, 2022.

[Marivate et al.] Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. Puoberta: Training and evaluation of a curated language model for setswana. *ArXiv*.

[Mburu, 2023] Nancy Njeri Mburu. Developing an e-learning module for paratransit and bus rapid transit (brt) drivers in the nairobi metropolitan area in kenya. 2023.

[Muhammad et al., 2023a] Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.862. URL https://aclanthology.org/2023.emnlp-main.862.

[Muhammad et al., 2023b] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. Afrisenti: A twitter sentiment analysis benchmark for african languages, 2023b.

[Muhammad et al., 2023c] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada, July 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.315. URL https://aclanthology.org/2023.semeval-1.315.

[Murçós, 2021] Francisco André Barreiros Murçós. Urban transport evaluation using knowledge extracted from social media. 2021.

[Mwaura, 2020] Naomi Njeri Mwaura. Making urban transport and public spaces safer for women. *THE JUST CITY*, page 101, 2020.

[Myoya et al., 2023] Rozina Lucy Myoya, Fiskani Banda, Vukosi Marivate, and Abiodun Modupe. Fine-tuning multilingual pretrained african language models. In *4th Workshop on African Natural Language Processing*, 2023.

[Nicholas and Bhatia, 2023] Gabriel Nicholas and Aliya Bhatia. Lost in translation: Large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*, 2023.

[OECD, 2023] OECD. Ai language models: Technological, socio-economic and policy considerations. *OECD*, 352:1, 2023.

[Ogueji et al., 2021] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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, 2021.

[Qi et al., 2020] Bing Qi, Aaron Costin, and Mengda Jia. A framework with efficient extraction and analysis of twitter data for evaluating public opinions on transportation services. *Travel behaviour and society*, 21:10–23, 2020.

[Qi et al., 2019] Zhaodi Qi, Yong Ma, and Mingliang Gu. A study on low-resource language identification. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1897–1902. IEEE, 2019.

[Regulator, 2021] Railway Safety Regulator. State of safety report 2020/21. Technical report, Railway Safety Regulator, 2021.

[Republic of South Africa, 1996] Republic of South Africa. Constitution of the republic of south africa. URL: https://www.justice.gov.za/constitution/SAConstitution-web-set.pdf, 1996. Accessed: 2023-11-05.

[Rieser-Schüssler and Axhausen, 2013] Nadine Rieser-Schüssler and Kay W Axhausen. Identifying chosen public transport connections from gps observations. In *TRB 92nd Annual Meeting Compendium of Papers*, pages 13–0588. Transportation Research Board, 2013.

[SABS, 2009] SABS. Railway safety management. Standard SANS 3000-1:2009, South African Bureau of Standards, Pretoria, South Africa, 2009.

[Salcianu et al., 2018] Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, et al. Compact language detector v3. 2018.

[Shmueli, 2019] Boaz Shmueli. Multi-class metrics made simple, part ii: the f1-score. https://medium.com/towards-data-science/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1, 2019. Accessed: 2023-04-04.

[Statista, 2023] Statista. The statistics portal for market data, market research and market studies, 2023. URL https://www.statista.com/.

[Tauchmann, 2021] Christopher Tauchmann. Advanced corpus annotation strategies for nlp. applications in automatic summarization and text classification. 2021.

[Vicente and Reis, 2016] Paula Vicente and Elizabeth Reis. Profiling public transport users through perceptions about public transport providers and satisfaction with the public transport service. *Public Transport*, 8:387–403, 2016.

[Wang et al., 2022] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 2022.

[Weights & Biases, Inc., 2023] Weights & Biases, Inc. Weights & biases. https://wandb.ai/site, 2023. Accessed: 2023-01-09.

[WHO, 2022] WHO. Kenyan government, world health organization, bloomberg philanthropies launch new initiative to reduce road crash deaths, 2022. URL https://www.who.int/news/item/25-05-2022-kenyan-government--world-health-organization--bloomberg-philanthropies-launch-n
Accessed: [2023/10/20].

Appendix A

Experimentation Logs

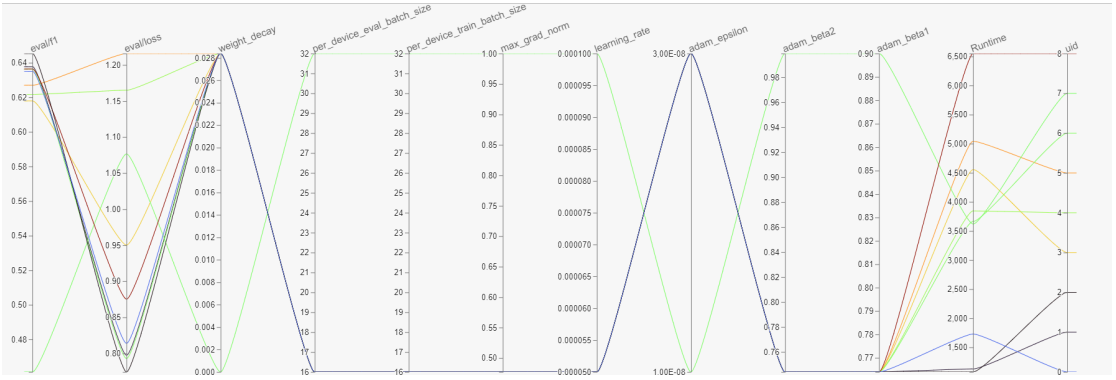

FIGURE A.1: AfriBERTa Experimentation Logs

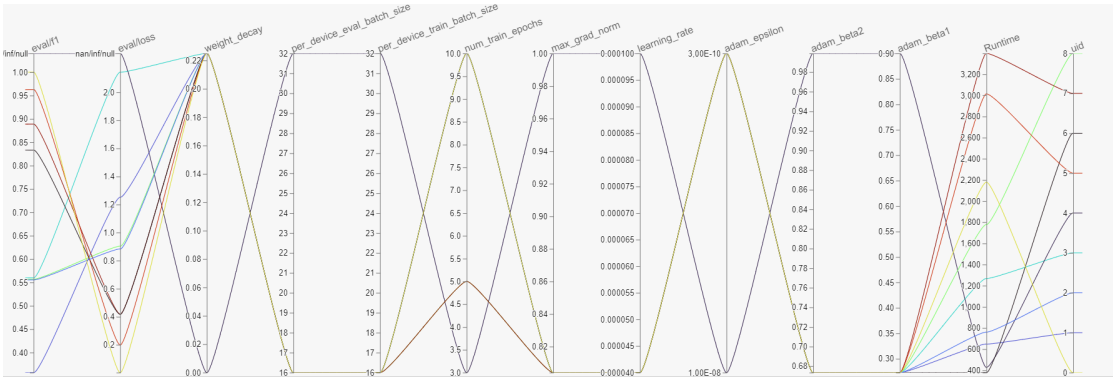

FIGURE A.2: AfroXLMR Experimentation Logs

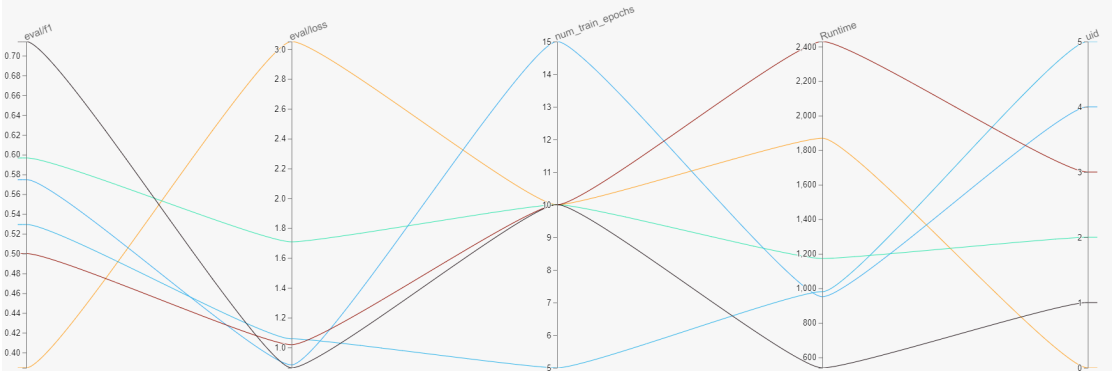

FIGURE A.3: AfroLM Experimentation Logs

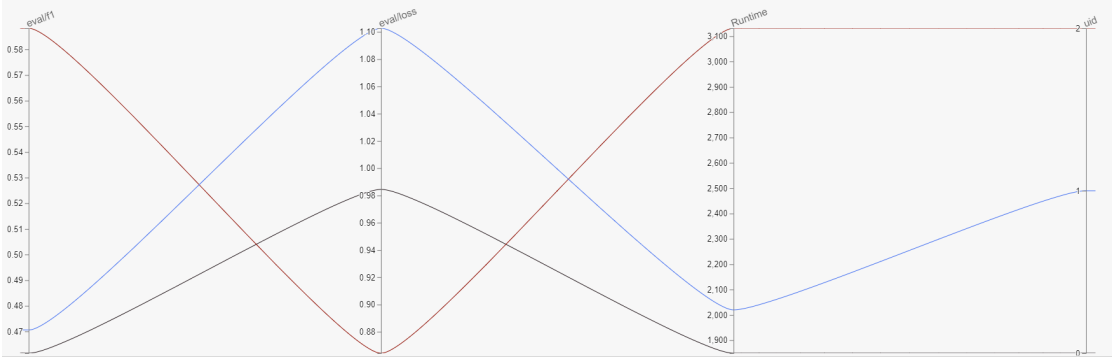

FIGURE A.4: PuoBERTa Experimentation Logs