# Developing a code-mixed sentiment analysis dataset of Xitsonga-English music review

*Blessing Nkuna*
*Computer Science Department, University of Limpopo, Sovenga 0727, South Africa*
*nkuna_b@outlook.com*

*Thipe I Modipa*
*Computer Science Department,*
*University of Limpopo, Sovenga 0727, South Africa*
*Center for Artificial Intelligence Research (CAIR), South Africa*
*thipe.modipa@ul.ac.za*

*Simon P. Ramalepe*
*Computer Science Department,*
*University of Limpopo, Sovenga 0727, South Africa*
*University of Limpopo*
*simon.ramalepe@ul.ac.za*

**Abstract**

Sentiment analysis is the process of classifying text emotions as positive, negative or neutral. Code-mixed sentiment analysis refers to the classification of text's sentiments that contains two or more languages. There are limited studies developed for sentiment analysis on South African code-mixed languages and this is due to the absence of annotated dataset. The purpose of the study was to collect code-mixed text data for the Xitsonga-English language pair. The study collected Xitsonga-English code-mixed comments for music reviews from a YouTube channel. After the data was collected, tokenization using a python library called natural language toolkit was performed. Subsequently, we analyzed the comments for the presence of code-mixing. The collected Xitsonga-English code-mixed data would be suitable to build a sentiment analysis model.

**Keywords:** Code-mixed, Sentiment analysis, Xitsonga-English language

## 1 Introduction

Xitsonga is one of the 12 official languages of South Africa, with a population of 2.2 million Xitsonga speakers (Alexander, 2023). In the South-East of Southern Africa, Xitsonga is dispersed over an extensive area. It is also widely spoken in Zimbabwe and southern Mozambique and is known as Xichangana. Speakers of this language tend to mix it with English language in their daily conversation or when they express their views (Zerbian, 2007).

Sentiment analysis is a natural language processing technique that monitors an organization's online discussions about their brand, product and service by extracting information from the content in the source material (Chakravarthi *et al.,* 2021). Organizations, historically, used sentiment analysis to determine customers' sentiments on products using product reviews. However, in recent years, sentiment analysis has expanded to include social media texts (Mäntylä *et al.,* 2018). Sentiment analysis can be performed on either monolingual reviews or code-mixed reviews.
A single sentence that combines two different languages is referred to as a code-mixed sentence (Chandu *et al.*, 2018). Code-mixing has grown in popularity among social media users to use more than one language. It has been more common for existing studies to analyze sentiments from monolingual data rather than mixed language text data (Srinivasan and Subalalitha, 2021). Monolingual data refers to data that contain a single language.

A study by Prabhu *et al.* (2013) shows that due to the absence of suitable annotated datasets, research conducted on code-mixed sentiment analysis has been minimal. In another study, Dutta *et al.* (2021) mentioned that sentiment analysis models for messages, tweets, or reviews with more than one language, such as English-Hindi, English-Chinese, English-Tamil, or Xitsonga-English, are less developed. The existing sentiment analysis models focus more on monolingual English language than code-mixed languages.

To the best of our knowledge, only a study by Muhammad *et al.* (2023) proposed a sentiment analysis model for Xitsonga-English code-mixed dataset. They developed a sentiment analysis model for the Xitsonga language that is spoken in Mozambique. In their study, the authors proposed a baseline experiment, which took into account three scenarios: (1) monolingual baseline models based on multilingual pre-trained language models for 12 AfriSenti languages with training data; (2) multilingual training of all 12 languages; and (3) zero-shot transfer from any of the 12 languages. They also developed a code-mixed sentiment analysis corpus for 13 other different African languages including Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, and Yorùbá. They used Twitter Academic API to collect the Twitter dataset. The purpose of the research study is to collect code-mixed text data for the Xitsonga-English language pair for the development of sentiment analysis model.

This paper is organized as follows, Section 2 discusses the related work of the study. In Section 3, we look at the methodology. Section 4, discusses the results and analysis. Lastly, conclusion follows in Section 5.

## 2 Related Work

As sentiment analysis advances, it faces the challenge of limited annotated code-mixed data (Choudhary *et al.,* 2018). The use of code-mixed languages in social media has expanded significantly due to its increasing popularity. Sentiment analysis classifiers for code-mixed data becomes a necessity since sentiment analysis models for monolingual data have been in existence and these classifiers cannot effectively be used for code-mixed data (Chakravarthi *et al.,* 2020).

Mabokela and Schlippe (2022) presented the first SAfriSenti corpus comprising English, Sepedi, and Setswana. Over forty thousand tweets of the three languages were gathered, with 36.6% being code-mixed. As a result of this work, they created sentiment lexicons for Sepedi and Setswana, which are incorporated into sentiment taggers through morphemes which indicate the class of sentiment. This study did not build any models for training and testing its dataset, nor did it evaluate its performance since no models were developed.

Mandal *et al.* (2018) conducted a study that focused on collecting a code-mixed corpus for sentiment analysis on Bengali-English languages. They collected 600 code-mixed sentences. The data was collected from Twitter using the public streaming API. Commonly used positive and negative Bengali words were used as keywords. They then filtered and cleaned the data and performed polarity classification. The polarities were validated using Bengali SentiWordNet. In order to annotate the languages and sentiment tags, they developed a system that helps in basic annotation. They developed two systems, one for language tagging and the other one for sentiment tagging; for language tagging they used a two-step modular by combining lexicon based modules along with a supervised learning module and for sentiment tagging they used a hybrid system for sentiment classification. They discovered that the majority of the tweets in both training and testing include relatively little neutral data. They also found out that the language tagger performed better than the sentiment tagger.

Sabri *et al.* (2021) collected, labeled and created a corpus for Persian-English code-mixed tweets. They then introduced a model which uses BERT pre-trained embeddings and translation models to automatically learn the polarity of the tweets. They collected 3640 tweets with labeled polarity, their tweets were collected using a Twitter API. They created a vectorized representation of the textual input in order to be able to fit the data into a machine learning model. To do so, they used these three steps: (1) Finding the non-Persian words in the sentence, (2) translated the non-Persian words using an automatic tool called Yandex, (3) created an embedding for the textual data by using the pre-trained multilingual BERT model. They fed the data into an ensemble model consisting of three Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and compared it with Naive

Bayes and random forest. The ensemble model outperformed the other two models.

Patra *et al.* (2018) presented a study on sentiment analysis of code-mixed data pairs of Hindi-English and Bengali-English collected from social media platforms. The Twitter API was used to collect both Bengali and Hindi code-mixed data from twitter. The baseline systems were developed by randomly assigning sentiment values; they used Glove word embeddings, TF-IDF scores of word n-grams, sentiment lexicon based model, Support Vector Machine, Naive Bayes, Document Term Matrix, Convolutional Neural Networks, Bi-LSTM. The evaluation metrics precision, recall and f-score were used. The discovery was that some deep learning models are successful for many NLP tasks and the n-gram based model had better results for F1-score.

Singh *et al.* (2018) presented a unique language tagged and POS-tagged dataset of code-mixed English-Hindi tweets. The study discussed a POS tagging model that was trained on the collected dataset. Twitter's streaming API to collect the tweets was used. For data annotation, two bilingual speakers fluent in English and Hindi annotated the tweets. The POS tagging model using Conditional Random Field (CRF) and LSTM Recurrent Neural Networks were developed. The findings showed that the CRF had the best performance.

Sabty *et al.* (2019) collected and developed the first annotated code-mixed for Arabic-English corpus. They further developed deep neural networks and word embedding for Arabic-English code-mixed text. The data used in the study was collected from three different sources: (1) The transcribed speech corpus through informal interviews, (2) Twitter using the streaming API, (3) Arabic ANERCorp data-set for Named Entity Recognition (NER) by translating some of the existing annotated Arabic NER data. They loaded two-word embedding models to cover the two languages since the corpus is multilingual. In their study, the Arabic word embedding model was generated and loaded into the NER system, it was then created and saved using the word2Vec (W2V) technique. The W2V algorithm used continuous Bag-of-Words, skip-gram, and deep learning in addition to these techniques. The W2V model was trained using an independent Arabic newswire dataset. The developed systems achieved a low F1-score because the training data contained only one of the two languages and the testing data contained mixed sentences.

Other code-switched corpora have been collected for South African languages (Ramalepe *et al.* 2022; Modipa *et al.* 2022). However, these corpora are not specifically suitable for sentiment analysis. The data was collected for speech recognition and text generation studies. The target languages were Sepedi and English. In this study we aim to develop Xitsonga-English code-mixed corpus. The developed dataset will be used to train a sentiment analysis model for South African Xitsonga language using Long short-term memory technique.

## 3 Methodology

In this section, we discuss the process used to collect Xitsonga-English code-mixed music comments from YouTube and how the data was cleaned and pre-processed. The study focused on code-mixed comments, that is, the comments that are a mix of Xitsonga and English.

### 3.1 Data Collection and Cleaning

The comments were collected from a Xitsonga YouTube channel, the channel posts Xitsonga songs, music festivals, and interviews with artists who are promoting their music. The comments were collected using a web scraping software. For every music post in the channel we copied the URL for that post and plugged it in the software and extracted the comments. To make sure that the comments were code-mixed, we manually went through the collected comments to confirm the presence of code-mixing. The study managed to collect 989 Xitsonga-English code-mixed music comments. The next step that followed after collecting the comments was to clean the comments; this included removing the numbers,

emojis, punctuations, stop-words and changing uppercase letters to lowercase.
Table 1 demonstrates the overall of how the data is structured. We show the comments after the cleaning process. There are 12 144 words which make 989 comments. Most comments contain 6 words; the data has an average comments of 12.28 which is approximately 12, the data also has a high standard deviation of 9.30 which means that the number of words in the comments are spread out from the mean. The comment with the most number of words has 85 words and the one with less has 2 words.

*Table 1: Summary Statistics for all the comments*

| | |
|---|---|
| Mean | 12.28 |
| Mode | 6 |
| Standard Deviation | 9.30 |
| Minimum | 2 |
| Maximum | 85 |
| Sum | 12144 |
| Count | 989 |

### 3.2 Tokenization

Table 2 shows a summary of the tokenized comments. It depicts the first 5 entries of the comments tokenized and the number of words each comment contains. The process of tokenization was done with the python library called Natural Language Toolkit (NLTK).

We observe that the number of English words in the first 5 comments is on average higher than the Xitsonga words. For example, in the first comments there are 8 English words compared to 2 Xitsonga words, and 5 English words compared to 1 Xitsonga word in the second last comment.

*Table 2: Tokenized comments*

| Comments | No of words |
|---|---|
| ['*thank', 'loads*', 'boti', '*helping', 'preserve', 'culture', 'keep', 'good', 'work',* 'inkomu'] | 10 |
| ['kulelo', 'mina', '*everything', 'song', 'reminds', 'boy', 'hood*', 'ka', 'bungeni'] | 9 |
| ['*trust', 'good', 'music', 'thanks', 'sharing*', 'dya', 'himetela'] | 7 |
| ['*made', 'miss', 'home', 'proud', 'say*', 'tsonga'] | 6 |
| ['*wow', 'hit*', 'na', 'hingisela', 'boti'] | 5 |

## 4 Result and Analysis

Table 3 shows the summary statistics of Xitsonga words in the collected dataset after the pre-processing step. There are 8 506 Xitsonga words in the dataset. We counted the number of words from each comment and found that the minimum number of Xitsonga words is 1 and the maximum number of Xitsonga words is 80. Most comments contain 4 Xitsonga words. The average number of Xitsonga words in the dataset is 8.6. These words have a high standard deviation of 8.19 which means that the number of Xitsonga words in the comments are more spread out from the average.

Table 4 shows the summary statistics of English words. There are 3 638 English words in the dataset. Our observation shows that there are comments with a single English word and multiple Xitsonga words. The maximum number of words in the comments is 34, and most comments contain 2 English words in them. The average number of English words is 3.68 with a standard deviation of 3.10 which means that the number of English words in the comments are more spread out from the average.

*Table 3: Summary Statistics for Xitsonga words*

| | |
|---|---|
| Mean | 8.60 |
| Mode | 4 |
| Standard Deviation | 8.19 |
| Minimum | 1 |
| Maximum | 80 |
| Sum | 8506 |
| Count | 989 |

Table 5 shows the frequency of Xitsonga and English words, the first entry has a column called number of words, with the first entry being 1 and a frequency of 166 for English and 18 for

Xitsonga. This observation suggests that there are 166 sentences that contain only a single word of English and 18 comments that have only 1 Xitsonga word in them. In a case of rows with zero, that will mean that there are zero sentences that contain a certain number of words for one of the languages, for example there are zero comments that contain 17 words of Xitsonga in them and there are zero comments that contain 39 English words in them.

Table 6 shows the frequency of each word in the dataset. It shows some part of the frequencies but not all of them. For instance, in every comment the study counted the number of Xitsonga and English words. We then accumulated the comment with the same number of words together, for example, with the first entry. There are 5 comments that contain only 2 words.

*Table 4: Summary Statistics for English words*

| | |
|---|---|
| Mean | 3.68 |
| Mode | 2 |
| Standard Deviation | 3.10 |
| Minimum | 1 |
| Maximum | 34 |
| Sum | 3638 |
| Count | 989 |

*Table 5: Frequency for Xitsonga and English words in a comment.*

| Number of words | English | Xitsonga |
|---|---|---|
| 1 | 166 | 18 |
| 2 | 233 | 174 |
| 10 | 10 | 0 |
| 17 | 1 | 0 |
| 23 | 1 | 11 |
| 34 | 1 | 7 |
| 39 | 0 | 5 |
| 44 | 0 | 1 |
| 51 | 0 | 1 |
| 67 | 0 | 1 |
| 74 | 0 | 1 |
| 80 | 0 | 1 |

*Table 6: Frequency for all the comments*

| Number of words | Number of Comments |
|---|---|
| 2 | 5 |
| 4 | 65 |
| 7 | 235 |
| 10 | 235 |
| 20 | 32 |
| 28 | 12 |
| 39 | 4 |
| 50 | 1 |
| 66 | 1 |
| 74 | 3 |
| 85 | 1 |

Figure 1 shows the number of comments versus number of words. The dataset is dominated by comments with 7 and 10 words in length. The longer comments are not that many. Again Figure 2 shows the frequency graph of both the Xitsonga and English words. The shorter words are dominating in the dataset for both the languages.

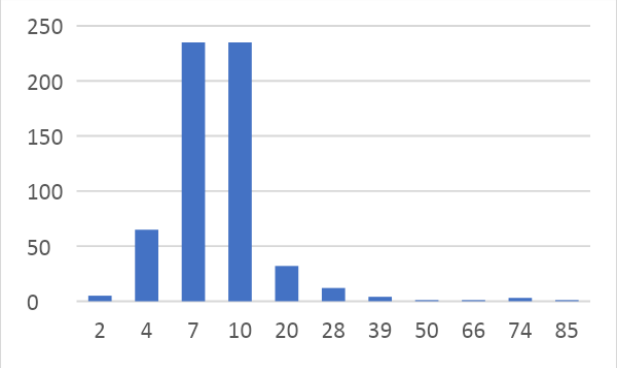


*Figure 1: Frequency graph of all the comments*

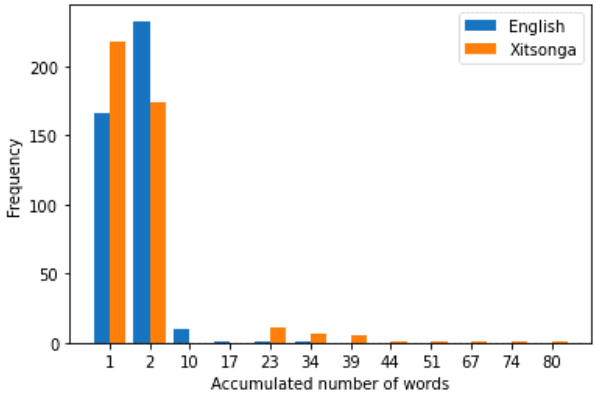


*Figure 2: Frequency graph of Xitsonga words and English words*

*Table 7: Top Xitsonga and English words.*

| Top Xitsonga | Occurrence | Top English | Occurrence |
|---|---|---|---|
| Ku (PP) | 199 | Music (N) | 70 |
| na (C) | 185 | President (N) | 65 |
| ya (PP) | 170 | Album (N) | 61 |
| ka (PP) | 135 | Love (V) | 55 |
| wena (P) | 118 | Song (V) | 52 |
| va (PP) | 102 | Best (N) | 50 |
| loko (C) | 100 | Keep (V) | 45 |
| wa (PP) | 94 | Good (A) | 39 |
| ni (C) | 87 | Hit (N) | 37 |
| mina (P) | 69 | Problem (N) | 35 |

*Key: P- Pronoun, N- Noun, V- Verb, C- Conjunction, PP- Preposition, A- Adjective.*

Table 7 shows the top 10 most used words in the dataset for both English and Xitsonga languages. For Xitsonga language, the most occurring words are prepositions whereas for the English language, the most occurring words are nouns. Also, the occurrence of the English words, is below 100 while for Xitsonga words the highest occurrence is almost 200.

## 5 Conclusion

The data that was collected is suitable to build a code-mixed sentiment analysis model for Xitsonga-English language pair. The average code-mixed ratio suggests that there are 9 words of Xitsonga in a 12-word comment and 3 English words in a 12-word comment; this shows that more users prefer using Xitsonga language more than English language in a code-mixed comment. For future work, the study will even collect monolingual comments for both English and Xitsonga. The study will encourage additional sentiment analysis research on South African code-mixed languages. The collected dataset is suitable for sentiment analysis because the existing research in this domain lacks sufficient code-mixed datasets.

## References

Alexander, M 2023, 'The 11 languages of South Africa' https://southafrica-info.com/arts-culture/11-languages-south-africa/

Chakravarthi, BR, Priyadharshini, R, Muralidaran, V, Suryawanshi, S, Jose, N, Sherly, E & McCrae, JP 2020, 'Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text', *ACM International Conference Proceeding Series*, pp. 21–24.

Chakravarthi, BR, Priyadharshini, R, Thavareesan, S, Chinnappa, D, Thenmozhi, D, Sherly, E, McCrae, JP, Hande, A, Ponnusamy, R, Banerjee, S & Vasantharajan, C 2021, 'Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text', *CEUR Workshop Proceedings*, 3159, pp. 872–886.

Chandu, KR, Loginova, E, Gupta, V, Van Genabith, J, Neumann, G, Chinnakotla, MK, Nyberg, E & Black, AW 2018 'Code-Mixed Question Answering Challenge: Crowd-sourcing Data and Techniques', *Proceedings of the Annual Meeting of the Association for Computational Linguistics,* (Cm), pp. 29–38.

Choudhary, N, Singh, R, Bindlish, I. & Shrivastava, M 2018, 'Sentiment Analysis of Code-Mixed Languages leveraging Resource Rich

Languages', *Computational Linguistics and Intelligent Text Processing,* pp 104-114.

Dutta, S, Agrawal, H & Roy, PK 2021, 'Sentiment Analysis on Multilingual Code-Mixed Kannada Language', *CEUR Workshop Proceedings*, 3159, pp. 908–918.

Mabokela, KR & Schlippe, T 2022, 'A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context', *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, pp. 70–77.

Mandal, S, Mahata, SK & Das, D 2018, 'Preparing Bengali-English Code-Mixed Corpus for Sentiment Analysis of Indian Languages', *Computation and Language* [Preprint].

Mäntylä, MV, Graziotin, D & Kuutila, M 2018, 'The evolution of sentiment analysis—A review of research topics, venues, and top cited papers', *Computer Science Review*, 27, pp. 16–32.

Modipa, T. I., & Davel, M. H. 2022. 'Two sepedi-english code-switched speech corpora'. *Language Resources and Evaluation*, 56(3), 703-727.

Muhammad, SH, Abdulmumin, I, Ayele, AA, Ousidhoum, N, Adelani, DI, Yimam, SM, Ahmad, SI, Beloucif, M, Mohammad, S, Ruder, S, Hourrane, S, Brazdil, P, Ali, FDMA, Davis, D, Osei, S, Bello, BS, Ibrahim, F, Gwadabe, F, Rutunda, S, Belay, T, Messelle, WB, Balcha, HB, Chala, SA, Gebremichael, HT, Opoku, B, & Arthur, S 2023, 'AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages', *Computation and Language* [Preprint].

Patra, BG, Das, D & Das, A 2018, 'Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL_Code-Mixed Shared Task @ICON-2017', *Computation and Language* [Preprint].

Prabhu, A, Joshi, A, Shrivastava, M & Varma, V 2013, 'Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text', *International Conference on Computational Linguistics,* pp. 2482-2491

Ramalepe, S., Modipa, T. I., & Davel, M. H. 2022. 'The Analysis of the Sepedi-English Code-switched Radio News Corpus'. *Journal of the Digital Humanities Association of Southern Africa*, 4(01).

Sabri, N, Edalat, A & Bahrak, B 2021, 'Sentiment Analysis of Persian-English Code-mixed Texts', *26th International Computer Conference, Computer Society of Iran, CSICC 2021* [Preprint].

Sabty, C, Elmahdy, M & Abdennadher, S 2019, 'Named Entity Recognition on Arabic-English Code-Mixed Data', *Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019*, pp. 93–97.

Singh, K, Sen, I & Kumaraguru, P 2018, 'A Twitter Corpus for Hindi-English Code Mixed POS Tagging', *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 12–17.

Srinivasan, R & Subalalitha, CN 2021, 'Sentimental analysis from imbalanced code-mixed data using machine learning approaches', *Distributed and Parallel Databases* [Preprint].

Zerbian, S 2007, 'A First Approach to Information Structuring in Xitsonga/ Xichangana', *SOAS Working Papers in Linguistics*, 15, pp. 65–78.