2024 REPORT

Building Multilingual Lexicons for Enhanced Sentiment Analysis and Translation in AI Systems

INF 791 ASSIGNMENT 3

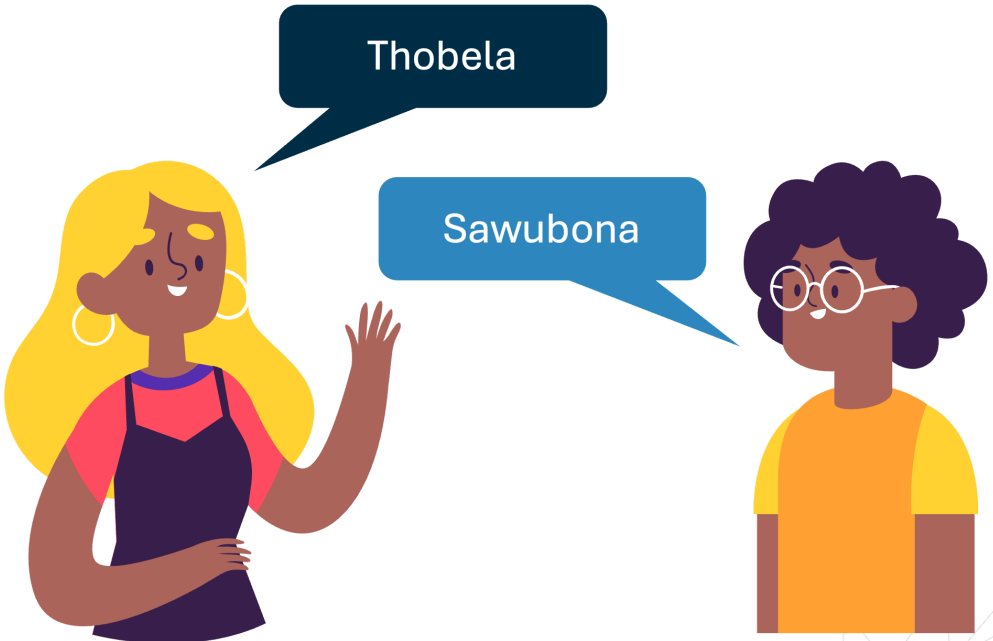


TABLE OF CONTENTS

i - TABLE OF FIGURES

List Of Images

ii - TABLE OF TABLES

List Of Tables

iii - PARTICIPANTS OF STUDY

Student Number								*Surname*	*Initials*
2	3	9	2	3	1	7	3	Malange	M
2	1	5	2	8	4	9	8	Mphahlele	M
2	1	5	5	5	3	4	7	Mphahlele	MM
2	1	5	5	9	5	1	2	Rampedi	MS
2	1	4	5	2	2	1	2	Mthembu	BKS
2	1	5	2	9	6	6	4	Puka	KB
2	0	4	9	6	7	0	3	Dibetso	KM
1	9	2	0	3	6	6	8	Mokhatla	PT
1	9	2	5	6	6	2	2	Shabangu	SP
1	9	2	2	1	8	8	7	Brand	CTJ
2	0	5	8	7	0	5	9	Skhonde	SC
2	1	7	2	3	6	4	9	Majikijela	S
2	1	6	1	7	5	9	8	Kubayi	B
2	1	6	9	1	3	9	9	Zwane	NM
2	0	4	4	9	1	3	6	Tebele	ME
2	0	5	3	4	9	5	8	Moeketsi	RM
1	9	3	6	3	9	6	7	Shozi	A
1	9	0	9	6	5	2	7	Nkuna	WN

Table 1. Table of Study Participant's Details

1 - INTRODUCTION

1.1 Problem Statement

South Africa's linguistic diversity presents challenges in sentiment analysis and translation across its 11 official languages. Existing lexicons often lack comprehensive coverage of local languages, impacting the accuracy of sentiment analysis in multilingual contexts.

1.2 Objectives

The primary objective of this report focuses on improving the accuracy and usability of AI systems in understanding sentiments expressed in diverse languages. In the report an expanded bilingual lexicon that supports sentiment analysis and translation between French, English, and South African languages were used to achieve the objective of this study.

The report focuses on utilising machine learning to compare the performance of not only the models using the different languages but also whether the corpus or the lexicon performed best. The following machine learning models were used: Support Vector Machine, Long Short Term Memory networks, Convolutional Neural Networks , Naive Bayes, Logistic Regression and Random Forest.

1.3 Scope

The scope of this study is to create a lexicon encompassing Ciluba , French and South African languages (e.g.,Sepedi, Afrikaans,Zulu) while addressing sentiment classification. The methodology includes data collection, processing, lexicon expansion, and the implementation of machine learning models for sentiment analysis.

2 - LITERATURE BACKGROUND

2.1 What is Sentiment Analysis?

Abdullah and Rusli (2021) define sentiment analysis as “a method or process of detecting and extracting a given subject such as opinion and attitudes from written and spoken language”. It is essentially the ability to detect the sentiment of a sentence or a topic and being able to rate its overall sentiment score and classify it as positive, negative, or neutral. It has gained more popularity especially with the advent of social media to determine the views that users have about certain topics that can include politics, certain brands and products or services to have a better-informed understanding of the consumer and their feelings and experiences. Social networking sites generate a large amount of data which cannot be processed or analysed by humans, this has led to the development of AI that can perform sentiment analysis.

However, sentiment analysis technology faces issues when it comes to deciphering multiple texts or opinions in multiple languages(Abdullah & Rusli, 2021). Multilingual sentiment analysis is aimed at improving the classification of text sentiment in multiple languages. The aim of this assignment is to build a lexicon that can accurately determine the sentiment of sentences in three different South African languages, namely Afrikaans, Sepedi and isiZulu.

2.2 Background and Justification

South Africa is a nation distinguished by its rich cultural diversity and multilingual society, with 11 official languages. Multilingual sentiment analysis is particularly valuable in this context as it allows the understanding of public opinion across different linguistic groups. In politics, it can guide policymaking and election strategies by revealing the sentiments of diverse communities. In healthcare, understanding public sentiment can enhance communication strategies, while in customer service, businesses can better cater to the needs of their multilingual customer bases. Moreover, keeping an eye on public opinion is crucial because it helps bring people together and tackle societal issues more effectively.

Focusing on low-resource languages like Sepedi and isiZulu is particularly significant due to their underrepresentation in digital resources (Mabokela & Schlippe, 2022). Developing sentiment analysis tools for these languages addresses the need for inclusive technological solutions, empowering communities and preserving linguistic heritage. By enhancing sentiment analysis capabilities for these languages, the project ensures that advancements in technology are accessible and beneficial to all, promoting a more equitable and inclusive digital landscape in South Africa.

2.2.1 Approaches

There are a number of approaches one can use to perform multilingual sentiment analysis; these include machine learning -based methods, parallel corpus-based methods, machine

translation-based methods and lexicon and corpus based methods (Araújo, Pereira & Benevenuto, 2020). The machine translation directly translates words from one language to another while preserving meaning, syntax and context. Machine learning methods train machine learning models using labelled data to classify and predict sentiment from text. Lexicon and corpus based use a lexicon to determine the overall sentiment in a sentence or written opinion where sentiment scores have been provided and the words in the sentence are matched to their sentiment score. The parallel corpus base has text and sentences from several languages, thereafter the paired texts are used to train models to map words and structures between languages. However, these four methods can be grouped into machine learning and lexical based approaches.

There are also different training approaches. These include sentence-based, and aspect-based approaches (Agüero-Torales, Abreu Salas & López-Herrera, 2021) as well as document level(ieee). These refer to the units by which the sentiment is analysed. The entire sentiment of a sentence is analysed for the sentence based. The specific aspects or features of an entity within a sentence are assessed in the aspect-based approach. For example, if a product is the subject of a discussion, the sentiment of that product is to be determined.

2.3 Key challenges

Producing the sentiment and sentiment score can be very subjective (Araújo et al., 2020) as it requires human beings in their own experiences and opinions to classify or label the words. Another challenge is finding individuals who are fluent in all the languages to be translated. There are a lot of homonyms (words that have different meanings but are spelt the same) which require a level of knowledge to be able to translate correctly.

Within each South African language, there are various dialects and regional variations. Capturing the nuances of each dialect in the sentiment analysis model is challenging due to the subtle differences in vocabulary and expressions. Additionally, detecting sarcasm and irony is inherently challenging in any language. This issue is exacerbated in low-resource languages, where the scarcity of annotated datasets that reflect these nuances is a critical concern (Mabokela & Schlippe, 2022).

In social media and everyday communication, it is common for speakers to alternate between languages within a single sentence or conversation. This practice is known as code-switching, this practice complicates sentiment analysis, as the model must accurately interpret multilingual text. The lexicon used in this assignment consists of 3000 words in Chiluba and French which were then translated into three South African languages. Although it is a step in the right direction, this is a very limited database. Data scarcity is one of the challenges of sentiment analysis. Some languages, like English, are very data rich and contain a large dataset of words while the others such as indigenous African languages, are not so popular and thus do not have a large dataset to work from. Most approaches were only developed for one the English language (Araújo et al., 2020), however communication on many platforms occurs in different languages. It is vital to test how the models perform using other languages as well as to adjust and even develop new models which are suited to different languages, and particularly South African, low-resource languages.

2.3.1 Strengths and weaknesses of lexicon based sentiment analysis

In previous studies such as the one by Dervenis, Kanakis and Fitsilis (2024), these machine learning based approaches performed better than the lexicon based approach when assessing the evaluation metrics. The other weaknesses include their struggle in handling the neutral sentiment class in addition to the positive and negative classes. It also struggles in recognising patterns within that class. Due to the dependence of the approach on the lexicon, it is very limited in its analysis as any word that appears outside the lexicon will not be recognised and taken into account into the sentiment analysis of the sentence. It is also very dependent on the quality of the lexicon; if the lexicon was built well, it will produce good results however if the lexicon was poorly built, the results will also be poor. Its inability to detect the context is a big issue as the sentiment scores cannot be adjusted to fit the context of the sentence. Due to not being able to detect contextual clues it can also not handle homonyms very well. It does have its advantages which includes its simplicity, as it directly maps the words and adds up the scores given for each word which translates into the second advantage which is its speed and efficiency as it does not require a computationally intensive process.

2.3.2 Lexicon based model

The sentiment analysis using the lexicon-based approach used rule-based logic. The steps in this approach include looking up the word in the lexicon to get the sentiment score of that word, then aggregating the scores of the words in the sentence to determine the total score of the entire sentence. If the score is close to zero, then it is classified as zero, if it is more than 0.05 it is positive and if it is less than -0.05 it is negative. The words and scores in the lexicon are used to construct the final score and sentiment which is then compared to the sentiment given by the human annotators.

2.4 Machine learning models

Supervised machine learning algorithms will be used for this task as the dataset used is labelled and has classified the words and sentences. The most popular models to use for sentiment analysis are Naïve Bayes, Support Vector Machine, CNN and RNN, LSTM, XAI. The models are trained on labelled datasets where the patterns of the words in the sentence are captured and the relationships between the words are learned to assess the overall sentiment in a sentence.

Preliminary objectives:

- Translate the French words into English and three South African languages (Zulu, Sepedi and Afrikaans)
- Clean the dataset by fixing the French words, nature and sentiment scores
- Use machine learning algorithms to perform the sentiment analysis
- Evaluate the performance of the sentiment analysis
- To assess which language performs better

2.5 Evaluation metrics

The standard and familiar evaluation metrics will be used to assess the performance of the machine learning methods. These include precision which measures how many predictions were correct, recall to measure how many positive or negative sentiments were correctly identified, the F1 score which is the balance between precision and recall, accuracy which measures the correct predictions as well as the ROC curve and AUC to measure the performance of binary classifiers and a confusion matrix to visually assess the type of errors and how the model performs in the different classes.

Evaluation Metrics

Standard evaluation metrics will assess the performance of the machine learning models:

- **Accuracy**: Measures the proportion of correct predictions.
- **Precision**: Assesses how many selected items are relevant.
- **Recall**: Evaluates how many relevant items are selected.
- **F1 Score**: Harmonic mean of precision and recall.
- **Confusion Matrix:** Visual representation of prediction errors.
- **ROC Curve and AUC**: Measure the performance of binary classifiers.

2.6 XAI in sentiment analysis

There is a desire to understand why the machine learning models classify sentiment in their various classes, this is where explainable AI (XAI) comes in. This method tries to uncover and understand what the machine learning model has learned during the training and how it applies that to predict a particular observation, in this case, the sentiment of a sentence (So, 2021). Deep learning models are “black boxes” because their inner workings are not known and regarded as a mystery, this method allows a light to be shined for transparency and for the correct improvements to be made to the model to increase the accuracy of its predictions. To secure rationality and give explanations for the basis of the decision making XAI is helpful. In this use case the XAI can map the feature importance to illustrate words which had the most influence in classifying the sentiment. For more complex sentences that go beyond the scope of purely positive or negative sentences but convey different types of emotions, subjectivity, intent, sarcasm, and topic being discussed. One effective XAI approach is LIME (Local Interpretable Model-agnostic Explanations), which helps illustrate feature importance by mapping which words had the most influence on the sentiment classification

3 - DATA COLLECTION AND PREPARATION

3.1 Dataset

The initial TSHIKAMA dataset contains three thousand French and Ciluba words that are labelled with sentimental scores. The dataset is expanded further by including translations from the English language and also three other South African languages. The chosen South African languages are as follows: Afrikaans, Sepedi and Zulu. The sentimental score represents either a positive sentiment or a negative sentiment. The positive sentimental scores range from 1 to 9, while the negative sentimental scores range from -1 to-9. In addition to that, the words are grouped according to their nature. The different natures featured for the words are namely: verbs, nouns, conjunctions, adverbs prepositions and Personal pronouns

3.2 Data Description

The data overview for the TSHIKAMA dataset, is as follows:

- **Ciluba**: Three thousand words in the Ciluba language
- **FRANCAIS**: Three thousand words in the French language
- **ENGLISH**: English translation of the three thousand words
- **Afrikaans**: Afrikaans translation of the three thousand words
- **Sepedi**: Sepedi translation of the three thousand words
- **Zulu**: Zulu translation of the three thousand words
- **Scores**: Sentiment scores that ranges from -9 to 9
- **Sentiment**: Identifies if the sentiment score is negative, neutral or positive
- **Nature**: The words are distinguished between verbs, nouns conjunctions, adverbs,prepositions and Personal pronouns

3.3 Pre-Processing

During the Pre-Processing stage, the EDA process is used. What is EDA? The full abbreviation of EDA is Exploratory Data Analysis which is defined as the exploration of data through various ways of data analysis. There are numerous reasons to substantiate the importance of implementing EDA.

What is the significance of implementing the EDA process on the TSHIKAMA dataset? Firstly, the implementation of the EDA process helps us understand the themes and patterns

of the dataset (Aasir, 2024). Secondly, the distribution of the sentimental values is analysed. Lastly, anomalies that need to be removed are detected from theTSHIKAMA dataset (Aasir, 2024).

(1) Missing values

The number of missing values are detected in all the columns in the TSHIKAMA dataset. A count of the number of missing values found in each column is provided. The findings indicate that the Sepedi, Afrikaans and Zulu columns have a number of missing words. As a result, the missing words are removed.

```
Missing Values in Each Column:
Ciluba        0
FRANCAIS      0
ENGLISH       0
Afrikaans     2
Sepedi        5
Zulu          1
SCORE         0
SENTIMENT     0
NATURE        0
dtype: int64
```

Figure 1. Results of missing values detected in each column

(2) Duplicates

The presence of duplicates can result in skewed insights. Removal of duplicates ensures consistency in the data. The dataset is examined for any duplicated rows. A Boolean analysis is done to analyse each row. As a result, False indicates that the row is unique and True indicates that the row is duplicated.

```
Duplicates:
 0       False
1        False
2        False
3        False
4        False
         ...
2995     False
2996     False
2997     False
2998     False
2999     False
Length: 3000, dtype: bool
```

Figure 2. Results after removal of duplicates

Measures are taken to drop in duplicates. Forty-nine words are removed from the total number of words in the TSHIKAMA dataset. Therefore, the total number of rows decreases to two thousand, nine hundred and fifty-four.

```
Shape after dropping duplicates (rows, columns): (2954, 9)
```

Figure 3. Total number of rows after duplicates were dropped

(3) Identifying mismatched sentimental scores

A comparison is done between the sentimental scores and the sentiment. Mismatches are detected between the two columns. The initial TSHIKAMA dataset consists of positive sentiments that are labelled with negative sentimental scores while the negative sentiments are labelled with positive sentimental scores. In the light of this, the sentiment is changed according to the sentimental score. A relation criteria is set for the negative sentiments and positive sentiments. For instance, any sentimental score that is greater than 0 is grouped into the positive sentimental category and any sentimental score that is less than 0 is grouped into the negative sentimental category. This relation criteria is used to make all the necessary changes to the TSHIKAMA dataset.

```
Rows with mismatched score and sentiment:
          Ciluba      FRANCAIS       English  Afrikaans     Sepedi      Zulu  \
351  disengelela  supplication  supplication  Smeekbede  Dikgopelo  ukunxusa
432      kunyema          fuir          flee      Vlugt     tshaba    baleka
503    kuimansha        jetter    throw away    weggooi      lahla     lahla

     SCORE SENTIMENT NATURE
351      3   Negatif    Mot
432      2   Negatif  Verbe
503      3   Negatif  Verbe
```

Figure 4. Results of rows with mismatched scores and sentiments

The relation criteria is implemented. As a result, all sentiments have been updated according to the sentimental value.

```
..   The sentiment values have been updated based on the score.
```

Figure 5. Confirmation message of updated sentiment values

The modules that were used for the EDA:

- Pandas
- Matplotlib
- Wordcloud
- Nltk(Natural Language Toolkit)

3.4 Data visualisation in the EDA

3.4.1 Word Cloud visualisation

A word cloud visualisation is used. The aim of the word cloud visualisation is to visualise negative words that are frequently mentioned (Suresh, 2020). The bigger the font, the more frequent the word appears (Suresh, 2020). Each language has a cloud visualisation that represents the frequency of the words that appear in that language.

3.4.1.1 Word Cloud for the Zulu Negative Sentiments

The word cloud visualisation for the Zulu Sentiment indicates that the key words that appear frequently are : “Ukudabuka”,”khala”, “amanga”,”Inzondo”, “ Hluleka”.

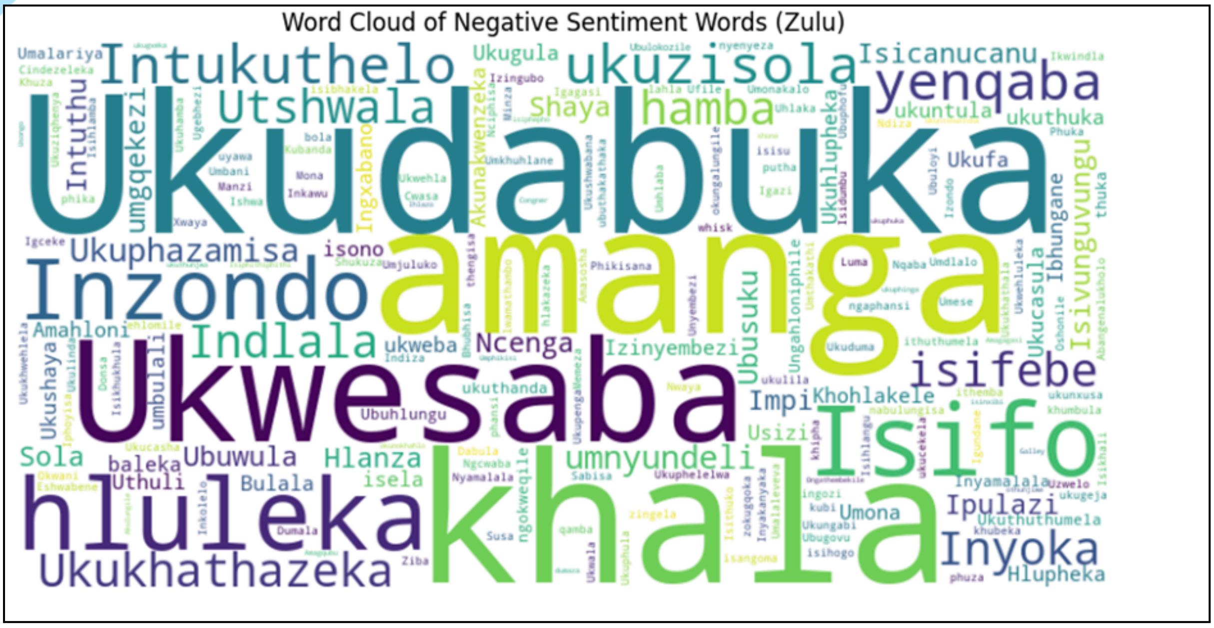


Figure 6. Word cloud representation of negative sentiment words in isiZulu

3.4.1.2 Word Cloud Visualisation for the Afrikaans Negative Sentiments

The word cloud visualisation for the Afrikaans Sentiment indicates that the key words that appear frequently is : *"Vrees, Weier, Huil ,Siekte, lieg, Hartseer, Haat"*

Figure 7. Word cloud representation of negative sentiment words in Afrikaans

3.4.1.3 Word Cloud Visualisation for the Sepedi Negative Sentiments

The word cloud visualisation for the Sepedi Sentiment indicates that the key words that appear frequently is *"go, maak, bolwetsi, tloga, Manyami, gana, ledimo, Hlkofetse".*

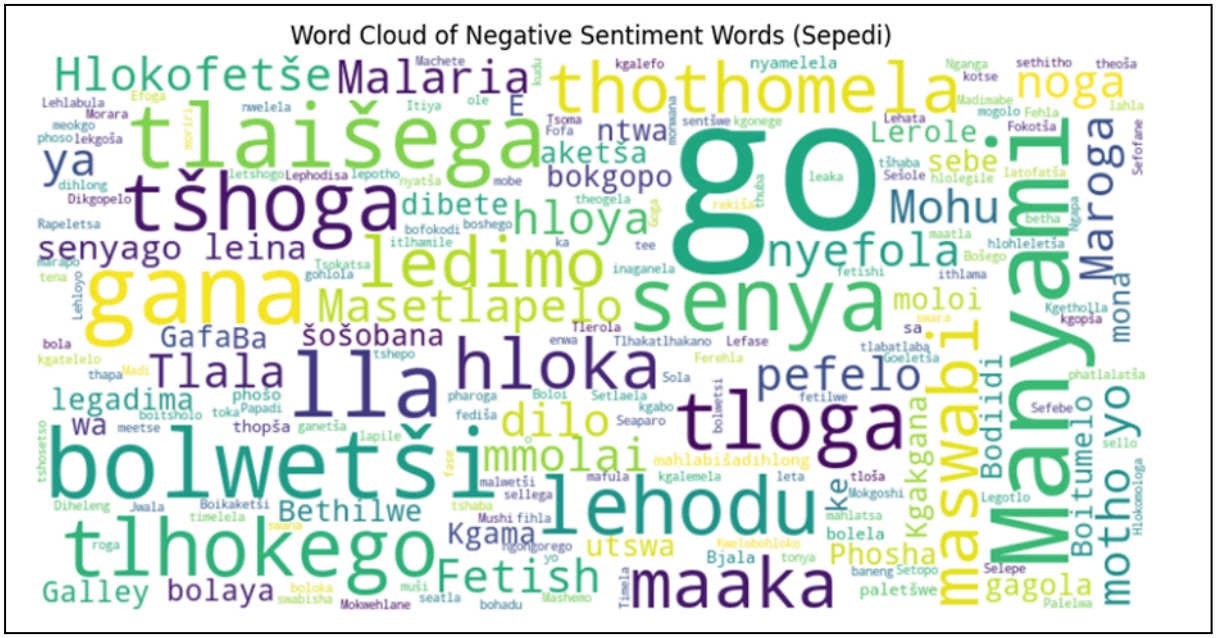


Figure 8. Word cloud representation of negative sentiment words in Sepedi

3.4.1.4 Word Cloud Visualisation for the English Negative Sentiments

The word cloud visualisation for the English Sentiment indicates that the key words that appear frequently are *"fear, Fetish, sadness, refuse, regret, lie, anger, leave."*

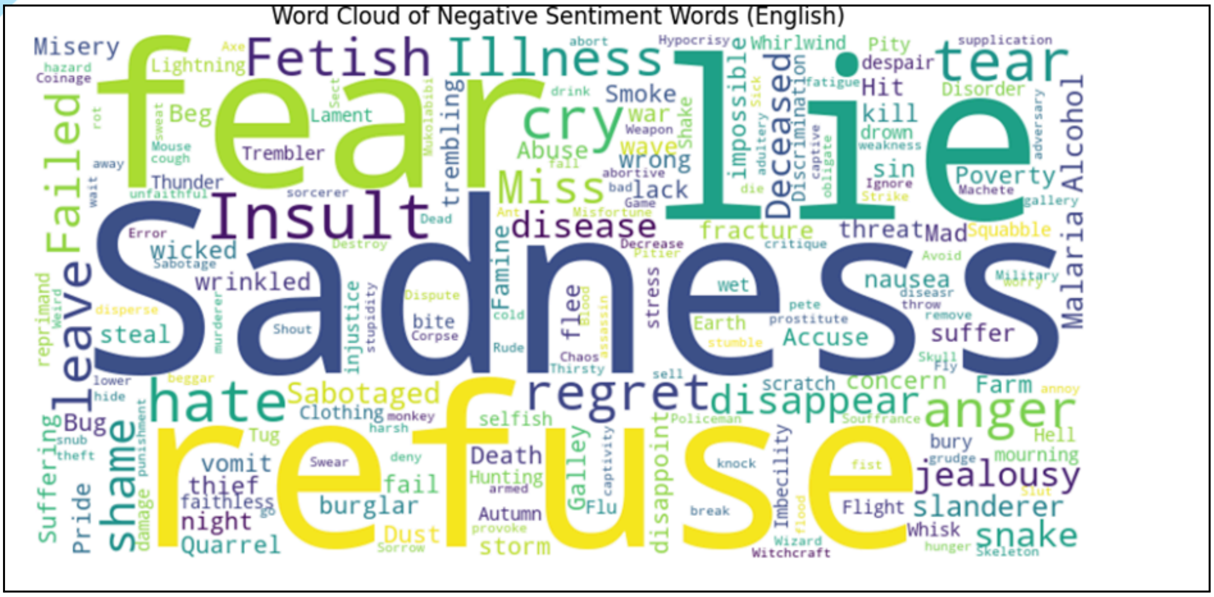


Figure 9. Word cloud representation of negative sentiment words in English

3.4.1.5 Word Cloud Visualisation for the French Negative Sentiments

The word cloud visualisation for the French Sentiments indicates that the key words that appear frequently are *"peur , tristesse, regret, Jalousie, mentir , partir , maladie , Larme".*

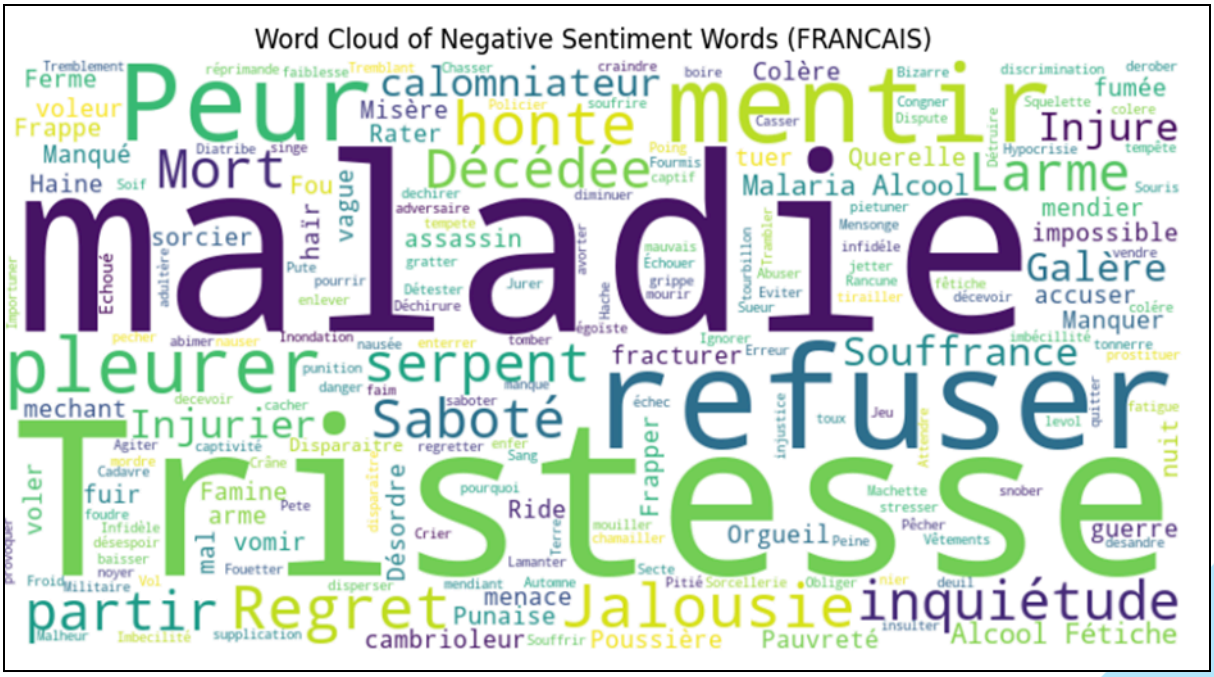


Figure 10. Word cloud representation of negative sentiment words in French

3.4.2 Distribution of words according to sentimental category

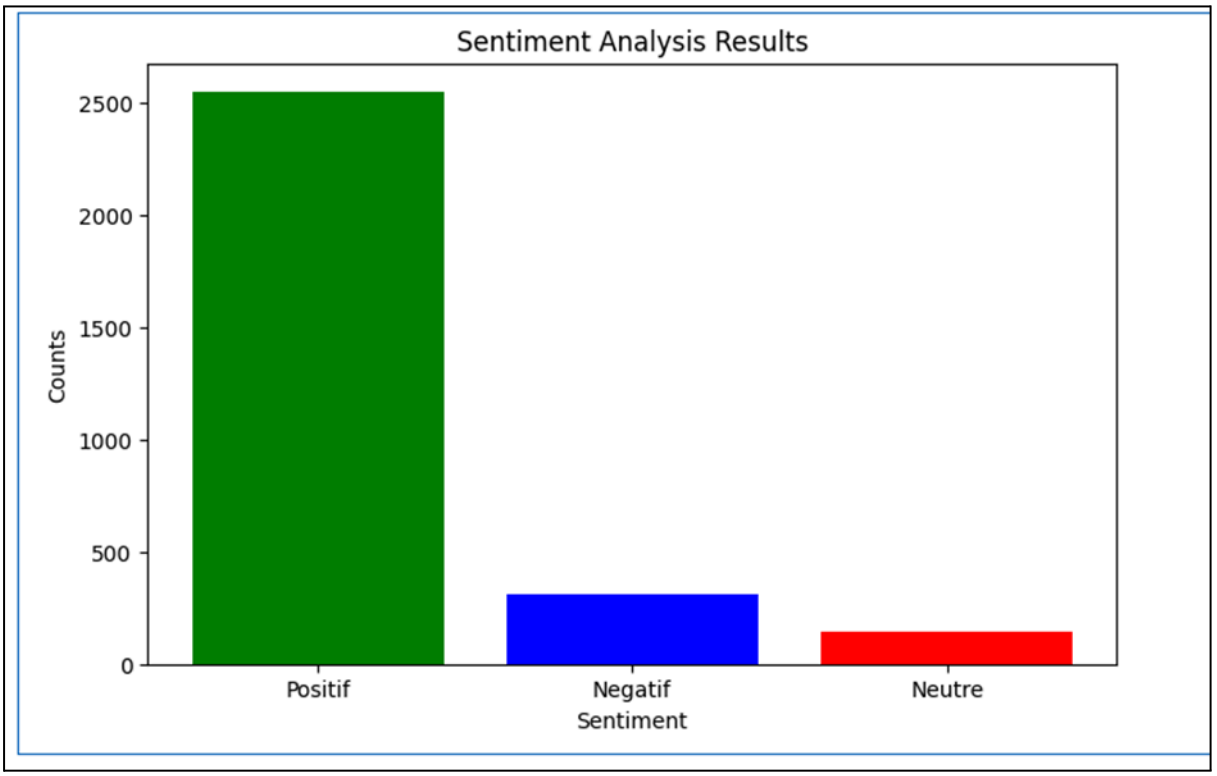


Figure 11. Bar graph of sentiment analysis results

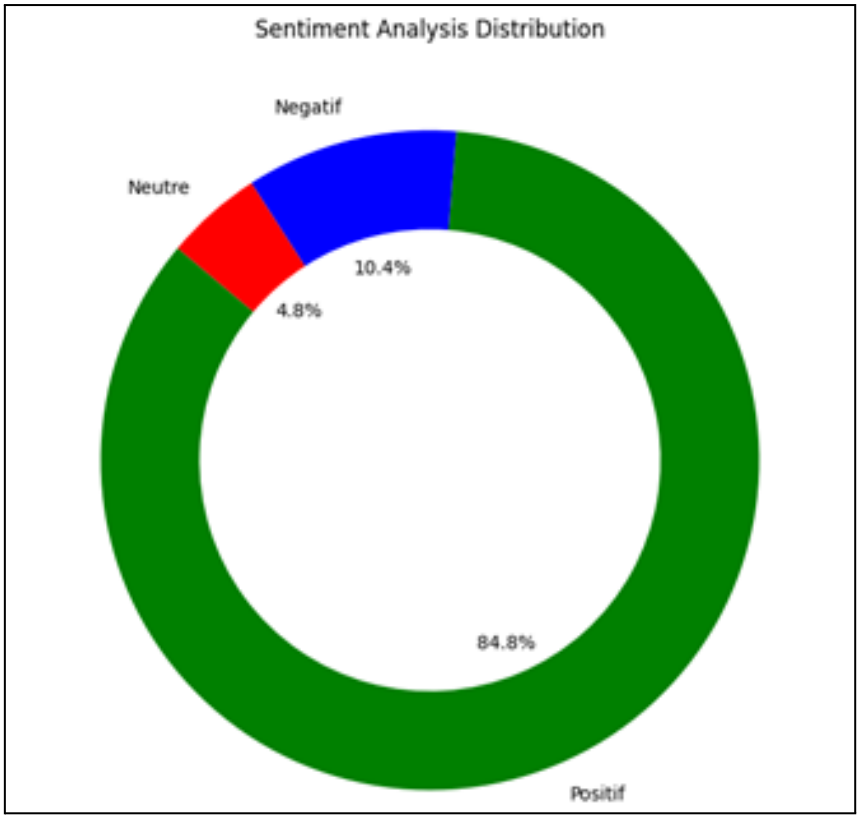


Figure 12. Donut graph representation of sentiment analysis distribution

Two visuals are used to visualise the distribution of the data. The visuals are namely the bar graph and the doughnut chart. The two charts display the word count of the distribution of the words according to the sentimental category.The bar graph and doughnut chart indicates that there are more positive sentiments than negative sentiments. In addition to that, some of the words in the dataset are not part of the negative or positive sentimental category. Therefore, the overall sentiments are positive

3.4.3 Distribution of Word Length by language

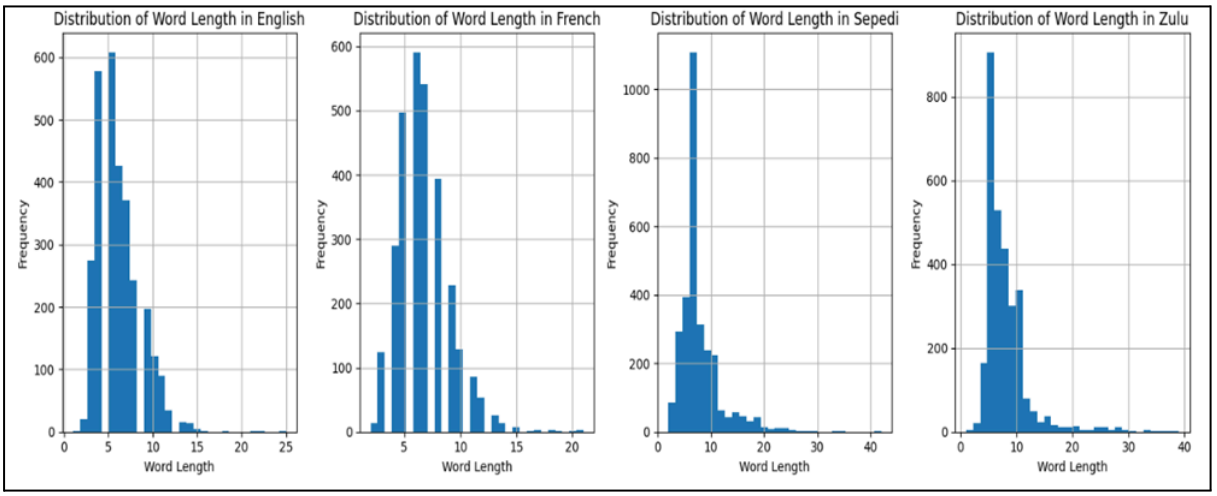


Figure 13. Bar graph representation of distribution of word length by languages

3.4.4 Bigram analysis

A bigram analysis is a type of n-gram analysis, it is a natural language processing technique that identifies the structure of words, the frequency and patterns in words or characters.

A higher count indicates that the bigram is more common or frequent in the text, while a lower count indicates it is less common. The bigram counts provide insights into the most frequently occurring two-word sequences in the text, which can be useful for various text analysis tasks, such as Identifying common phrases or collocations: The high-frequency bigrams can reveal the most common word combinations in the text Understanding language patterns: The bigram counts can help identify the typical word associations and structures used in the text. Comparing language usage: By comparing the top bigrams across different languages or text samples, you can identify similarities and differences in language usage. The bigram count information is valuable for understanding the lexical characteristics and patterns within the analysed text.

The figure below shows the top 20 bigrams identified in the English language. What can be observed is it identifies some bigrams that make sense such as “to walk”, “sit down” and others are not as precise or comprehendible such as “exchange branch”, “drinker calm”.

```
Top 20 Bigrams in ENGLISH:
seventy one: 9
one seventy: 8
to walk: 5
bush drinker: 5
drinker calm: 5
give birth: 4
sit down: 4
ardor aspect: 4
stock exchange: 4
exchange branch: 4
sadness anger: 3
to grow: 3
wake up: 3
to sit: 3
to do: 3
sheet metal: 3
To go: 3
To kiss: 3
Malaria Alcohol: 3
Money Mother: 3
```

Figure 14. Top 20 Bigrams in English

By comparing the top bigrams across different languages or text samples, you can identify similarities and differences in language usage.

The following are the top 20 bigrams for French, Sepedi and Zulu:

```
Top 20 Bigrams in FRANCAIS:
septante-un septante-un: 8
bourse branche: 5
branche bruit: 5
bruit buisson: 5
buisson buveur: 5
buveur calme: 5
Veine Ventre: 4
bourgeoise bourse: 4
ardeur aspect: 4
louer adorer: 3
Artère Barbe: 3
Cheveu Cheville: 3
Cil Cœur: 3
Epaule Estomac: 3
Gorge Hanche: 3
Intestin Jambe: 3
Jambe Joue: 3
Langue Larme: 3
Larme Lèvre: 3
Main Menton: 3
```

```
Top 20 Bigrams in Sepedi:
ya go: 21
masomesupa-tee masomesupa-tee: 8
go sepela: 7
go swana: 7
go swara: 6
go kgahlwa: 6
go dira: 5
o mogolo: 5
swana le: 5
sethokgwa monwi: 5
monwi homola: 5
na le: 5
dula fase: 4
tša go: 4
go thelela: 4
go ngwala: 4
ya tšhika: 4
Go tlaišega: 4
hlompha go: 4
molaodi wa: 4
```

```
Top 20 Bigrams in Zulu:
nanye amashumi: 14
amashumi amabili: 10
amashumi amathathu: 10
amashumi amane: 10
amashumi amahlanu: 10
amashumi ayisithupha: 10
amashumi ayisikhombisa: 10
ayisikhombisa nanye: 10
nesishiyagalolunye amashumi: 6
ukushintshaniswa kwe-sotck: 5
kwe-sotck igatsha: 5
igatsha umsindo: 5
umsindo ihlathi: 5
nambili amashumi: 5
nantathu amashumi: 5
nane amashumi: 5
nanhlanu amashumi: 5
nesithupha amashumi: 5
nesikhombisa amashumi: 5
nesishiyagalombili amashumi: 5
```

Figure 17. Top 20 Bigrams in French, Sepedi and isiZulu

3.4.5 Lexical diversity calculation for the different languages

```
Lexical Diversity in ENGLISH: 0.58
Lexical Diversity in FRANCAIS: 0.66
Lexical Diversity in Sepedi: 0.51
Lexical Diversity in Zulu: 0.60
```

Figure 18. Lexical Diversity in English, French, Sepedi and isiZulu

The lexical diversity value ranges from 0 to 1 (or 0% to 100%), where a higher value indicates a more diverse vocabulary, and a lower value suggests a more repetitive or limited vocabulary. A lexical diversity of 0.75 (or 75%) means that the text contains 75% unique words out of the total number of words. In other words, 75% of the words in the text are unique, and the remaining 25% are repeated. Lexical diversity is a useful metric for understanding the richness and diversity of the vocabulary used in a text. It can provide insights into the language proficiency, writing style, and the range of vocabulary employed in the analysed text.

4 - METHODOLOGY

We began our lexicon expansion process by translating the French words from the provided Tshikama lexicon Natural Language Processing (NLP) script into English. This was followed by translating the English words into Afrikaans, isiZulu, and Sepedi. This multi-stage expansion allowed our lexicon to perform sentiment analysis and translation across several languages. After completing these translations, we created new columns for each language, where each word was assigned a corresponding sentiment score (0 for neutral, 1 to 9 for positive, and -1 to -9 for negative words).

In this process, we encountered significant challenges due to the subtle differences in sentiment expression across languages. Many terms lacked direct translations in some African languages, and certain words carried varying connotations, complicating the task of maintaining appropriate sentiment across translations. These issues are echoed in existing literature, where the complexity of sentiment classification is noted, particularly for low-resource languages (Mabokela & Schlippe, 2022).

For sentiment computation, we utilised the sentiment classifications from our expanded lexicon to categorise terms into negative, positive, and neutral sentiments. Each entry in our dataset was assigned a score indicating its intensity, ranging from -9 to +9. We computed the overall sentiment for a sentence by summing the individual word scores, where a negative total indicated negative sentiment, a positive total indicated positive sentiment, and a near-zero total illustrated a neutral sentiment.

To ensure quality and consistency in translations, we employed Google Translate while paying careful attention to the sentiment of each word. This approach aligns with the principles of machine translation, which emphasises preserving meaning, syntax, and context (Araújo, Pereira & Benevenuto, 2020). Our focus on maintaining the appropriate tone helped ensure the translations accurately reflected the original meanings across languages.

We tested a variety of machine learning models for sentiment classification, including Random Forest, Logistic Regression, and Support Vector Machines (SVM). These models were trained on labelled sentiment data derived from our expanded lexicon. Evaluation metrics such as precision, accuracy, and recall were employed to assess model performance. This process also reflects contemporary approaches to multilingual sentiment analysis, which can be broadly categorised into machine learning-based methods and lexicon-based methods (Araújo, Pereira & Benevenuto, 2020).

While lexicon-based methods offer advantages in speed and efficiency, they often struggle with handling neutral sentiment and recognizing patterns within that class, as identified by Dervenis, Kanakis, and Fitsilis (2024). These limitations, combined with the challenges of data scarcity in many South African languages, necessitate the development of models

tailored to these low-resource languages. Additionally, the presence of homonyms and the variability of dialects add layers of complexity to sentiment analysis, echoing the challenges noted in previous research (Araújo et al., 2020).

In addressing these challenges, we ensured our lexicon comprised a robust dataset by integrating 3,000 words in Chiluba and French, translated into Afrikaans, isiZulu, and Sepedi. While this represents a step forward, it is crucial to acknowledge the limitations of this dataset in capturing the rich linguistic diversity present in South Africa. Our methodology emphasises the importance of context in sentiment analysis, particularly when dealing with complex sentences that may convey mixed emotions or sarcasm.

To evaluate the performance of our sentiment analysis, we applied standard evaluation metrics:

- **Accuracy:** Measures the proportion of correct predictions.
- **Precision:** Assesses how many selected items are relevant.
- **Recall:** Evaluates how many relevant items are selected.
- **F1 Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Visual representation of prediction errors.
- **ROC Curve and AUC:** Measures the performance of binary classifiers.

In pursuit of transparency and improved accuracy in our sentiment classifications, we also integrated Explainable AI (XAI) techniques. These techniques aim to illuminate the decision-making processes of our machine learning models, helping us understand the factors influencing sentiment predictions (So, 2021). By mapping feature importance, we can identify which words significantly impact sentiment classifications, particularly in complex scenarios involving sarcasm or mixed sentiments.

5 - RESULTS

5.1 Model Performance

This section looks at how well the machine learning models perform in this study, specifically their ability to classify sentiments from two different datasets: the expanded Tshikama dataset and the new corpus dataset. The performance metrics consist of accuracy, precision, recall, and F1-score, which are important for figuring out how effectively each model can predict sentiment categories.

5.1.1 Machine Learning Performance (Using the expanded Tshikama dataset)

The SVM, Naïve Bayes, Random Forest were all affected by the class imbalance of the nature column and sentiment column; thus, they all did not perform satisfactorily and gave very low accuracies. To further improve their performances the word nature classes can be balanced by not having too much of one class.

5.1.1.1 SVM

The SVM achieved an accuracy of 66.5 % which is an indication of problems with the training. The model is not performing well for all classes except classes 4, 5 and 6. The precision, recall, and F-1 score of the classes are 0.00. It is focused on classes 4, 5 and 6, overlooking the rest. The model is not predicting the classes correctly because class 5 has the most count in the dataset while class 4 has one count. The model is likely focusing on class 5 due to its high count resulting in low precision and recall for other classes. Balancing the class weight was used to improve the model but the accuracy went down. The model still focused on classes 4, 5, and 6, see fig 19-20 for the evaluation metrics.

```
Accuracy of SVM: 0.665
Classification report of SVM:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         6
           2       0.00      0.00      0.00         1
           3       0.00      0.00      0.00         1
           4       0.67      0.99      0.80       383
           5       0.53      1.00      0.69        18
           6       0.00      0.00      0.00         1
           8       0.00      0.00      0.00       187
           9       0.00      0.00      0.00         2
          10       0.00      0.00      0.00         1

    accuracy                           0.67       600
   macro avg       0.13      0.22      0.17       600
weighted avg       0.45      0.67      0.53       600
```

Figure 19. Evaluation metrics of SVM

```
Accuracy of SVM: 0.5766666666666667
Classification report of SVM:
              precision    recall  f1-score   support

           0       0.03      0.33      0.06         6
           1       0.00      0.00      0.00         0
           2       0.00      0.00      0.00         1
           3       0.00      0.00      0.00         1
           4       0.69      0.85      0.76       383
           5       0.41      1.00      0.58        18
           6       0.00      0.00      0.00         1
           7       0.00      0.00      0.00         0
           8       0.29      0.01      0.02       187
           9       0.00      0.00      0.00         2
          10       0.00      0.00      0.00         1

    accuracy                           0.58       600
   macro avg       0.13      0.20      0.13       600
weighted avg       0.54      0.58      0.51       600
```

Figure 20. Improved evaluation metrics of SVM

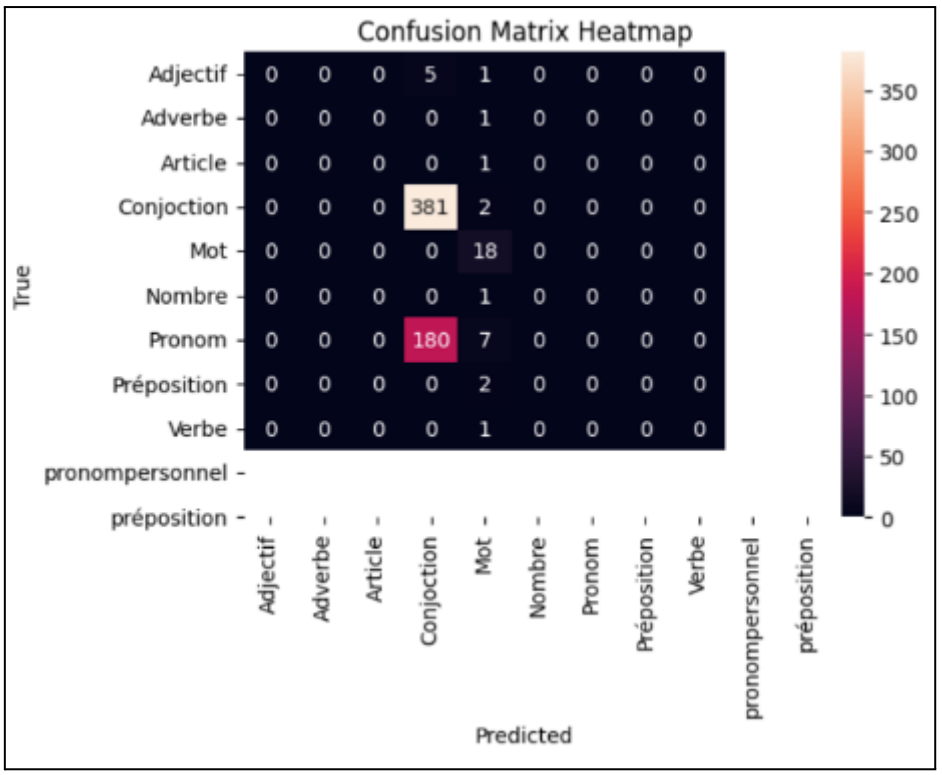


Figure 21. Confusion matrix heatmap of SVM

5.1.1.2 Naïve Bayes

As illustrated in the below, the accuracy of the model is 62.7% which highlights that the model did not perform exceptionally well, the low performance may have been a result of the low F1-score which indicates that the model is making a lot of false predictions thus leading to a dissatisfactory accuracy score (see Fig. 22). A number could be causing the low accuracy such as poor-quality data that isn't properly preprocessed or an imbalance in the classes which is something we can notice in this dataset. The classes are not equally distributed, some classes have more instances than others as shown in the classification report (see Fig. 23). This approach performed poorly, and a few things were added to help the model perform better and improve its accuracy (see Fig 24 & 25)

```
Accuracy: 0.627
Precision: 0.58
Recall: 0.63
F1 Score: 0.57
```

Figure 22. Evaluation metrics of Naïve Bayes

Classification Report:

	precision	recall	f1-score	support
Adjectif	0.00	0.00	0.00	6
Article	0.00	0.00	0.00	1
Conjoction	0.00	0.00	0.00	1
Mot	0.69	0.87	0.77	383
Nombre	0.53	1.00	0.69	18
Pronom	0.00	0.00	0.00	1
Verbe	0.41	0.13	0.20	187
pronompersonnel	0.00	0.00	0.00	2
préposition	0.00	0.00	0.00	1
micro avg	0.64	0.63	0.63	600
macro avg	0.18	0.22	0.18	600
weighted avg	0.58	0.63	0.57	600

Figure 23. Classification report of Naïve Bayes

The below figures 22 and 23 highlight the improved version of the Gaussian Naïve Bayes model. The improvement was conducted by performing hyperparameter tuning on the model using GridSearchCV. The cross-validation folds parameter was set to 10, to minimise the chance of the model overfitting and/or underfitting and therefore improving the overall performance of the model. As illustrated below, the overall performance of the model improved because of using GridSearchCV. (See figure 24 & 25)

```
Accuracy: 0.638
Precision: 0.41
Recall: 0.64
F1 Score: 0.50
```

Figure 24. Improved evaluation metrics of Naïve Bayes

Classification Report:

	precision	recall	f1-score	support
Adjectif	0.00	0.00	0.00	6
Article	0.00	0.00	0.00	1
Conjoction	0.00	0.00	0.00	1
Mot	0.64	1.00	0.78	383
Nombre	0.00	0.00	0.00	18
Pronom	0.00	0.00	0.00	1
Verbe	0.00	0.00	0.00	187
pronompersonnel	0.00	0.00	0.00	2
préposition	0.00	0.00	0.00	1
accuracy			0.64	600
macro avg	0.07	0.11	0.09	600
weighted avg	0.41	0.64	0.50	600

Figure 25. Improved classification report of Naïve Bayes

5.1.1.3 Random Forest

The Random Forest was used to predict the sentiment. The Fig 26 shows the results of the metrics which indicates the performance of the model. Fig 26 shows us that the accuracy for the Random Forest was 65% indicating that out of 100 predictions it means that 65% predictions were correct. There were 66% positive predictions made which are indicated by precision value. The recall shows that 89% of the predictions were true indicating that there are higher chances that the predictions made by the model are correct.

The classification report shows that precision shows that 0.66 of the predictions were correct. The sentiment was predicted correctly, and 0.57 of the cases were incorrectly predicted. The recall value is 0.89 which is high recall indicating that most of the sentiment value was correctly classified, it was a positive case. The f1-score is high which means the model which means that the precision and recall is more balanced. The support is higher for positive cases indicating that the model is likely to be reliable. The support shows that 562 of the sentiment of the Afrikaans word was predicted correctly and the 338 of the Afrikaans sentiment were predicted incorrectly.

In the confusion matrix shows that the model predicted 499 instances in positive were correctly predicted. In 319 (256+63) instances there were false positives. 82 out of the instances showed they were truly not predictive correctly.

Fig 28 illustrates the ROC curve that the model is likely to correctly predict as the line is closer to the diagonal line.

```
Accuracy: 0.65
Precision: 0.66
Recall: 0.89

Classification Report:
              precision    recall  f1-score   support

Not Positive       0.57      0.24      0.34       338
    Positive       0.66      0.89      0.76       562

    accuracy                           0.65       900
   macro avg       0.61      0.57      0.55       900
weighted avg       0.63      0.65      0.60       900
```

Figure 26. Classification report of Random Forest

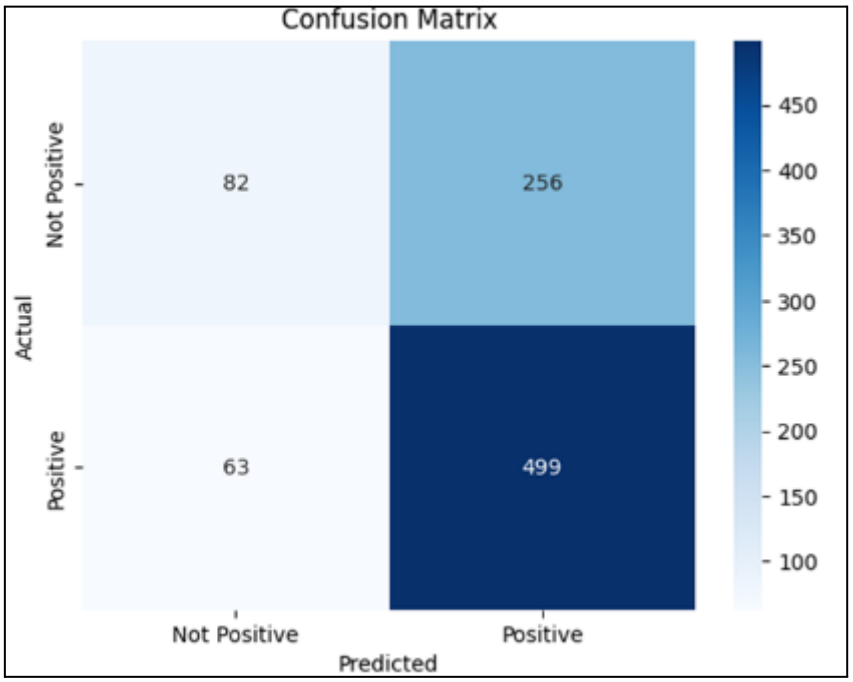


Figure 27. Confusion matrix of Random Forest

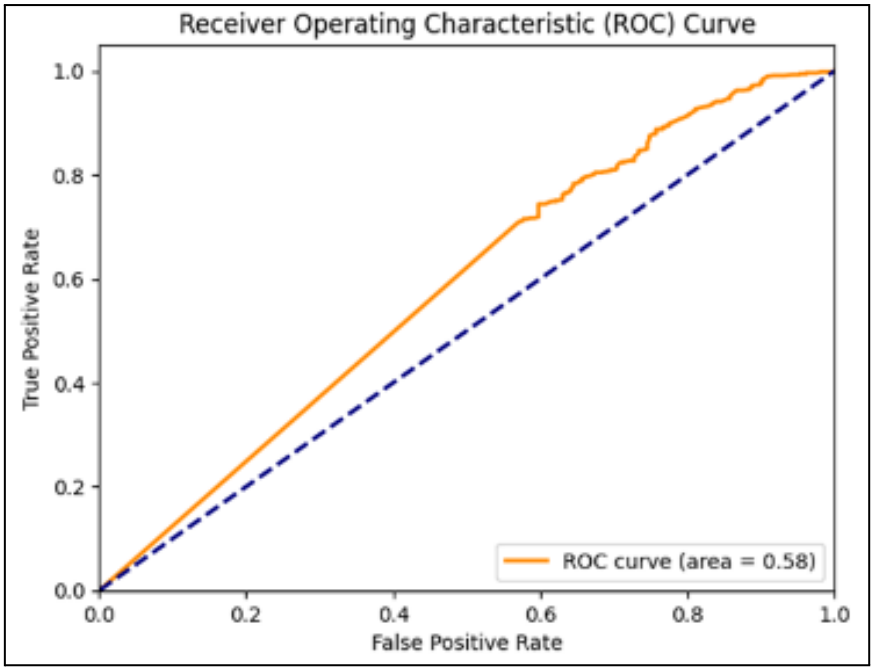


Figure 28. ROC Curve of Random Forest

5.1.1.4 Logistic Regression

For the logistic regression, the accuracy was 64 % indicating that out of 100 predictions it means that 64% predictions were correct. There were 0.64 positive predictions made which are indicated by precision value. The recall shows that 98% of the predictions were true, that the sentiment values were predicted correctly. The recall value should show that the positive instances are truly correct, and most cases were correct.

The classification report (Fig 29) shows that precision shows that 0.64 of the predictions were correct. The sentiment was predicted correctly, and 0.73 of the cases were incorrectly predicted. These suggestions that model is likely to make incorrect predictions. The recall value is 0.98 which is high recall indicating that most of the sentiment value was correctly classified, it was a positive case. The f1-score is high which means the model which means that the precision and recall is more balanced. The support is higher for positive cases indicating that the model is likely to be reliable. The support shows that 562 of the sentiment of the Afrikaans word was predicted correctly and the 338 of the Afrikaans sentiment were predicted incorrectly.

In the confusion matrix (Fig 30) shows that the model predicted 549 instances in positive were correctly predicted. In 316 (303+13) instances there were false positives. 35 out of the instances showed they were truly not predictive correctly.

The ROC curve shows that most of the predictions were true and correct.

```
Accuracy: 0.65
Precision: 0.64
Recall: 0.98
F1 Score: 0.78

Classification Report:
              precision    recall  f1-score   support

Not Positive       0.73      0.10      0.18       338
    Positive       0.64      0.98      0.78       562

    accuracy                           0.65       900
   macro avg       0.69      0.54      0.48       900
weighted avg       0.68      0.65      0.55       900
```

Figure 29. Classification report of Logistic Regression

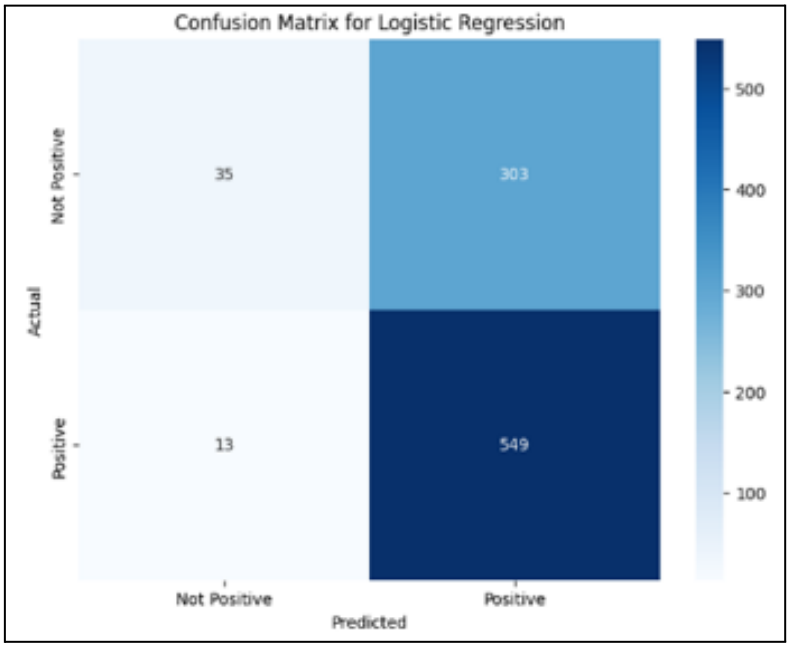


Figure 30. Confusion matrix of Logistic Regression

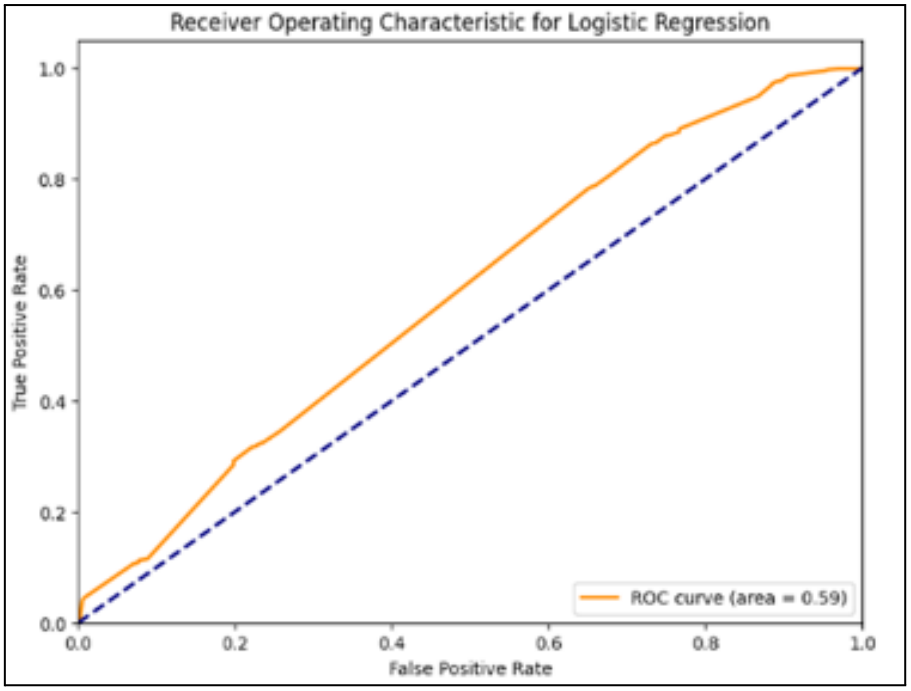


Figure 31. ROC Curve of Logistic Regression

5.1.2 Machine Learning performance (Using the newly created corpus dataset)

5.1.2.1 SVM

Afrikaans

The SVM model achieved an accuracy of 0.75 for Afrikaans, indicating that it performed relatively well. The precision, recall, and F1-score were strong for classes 0 and 1, indicating that the model could effectively classify most sentiments in Afrikaans. The F1-score for class 2 was 0.46, reflecting difficulties with this specific class. This could indicate a lack of sufficient data for this class or the complexity of understanding nuances in the sentiment.

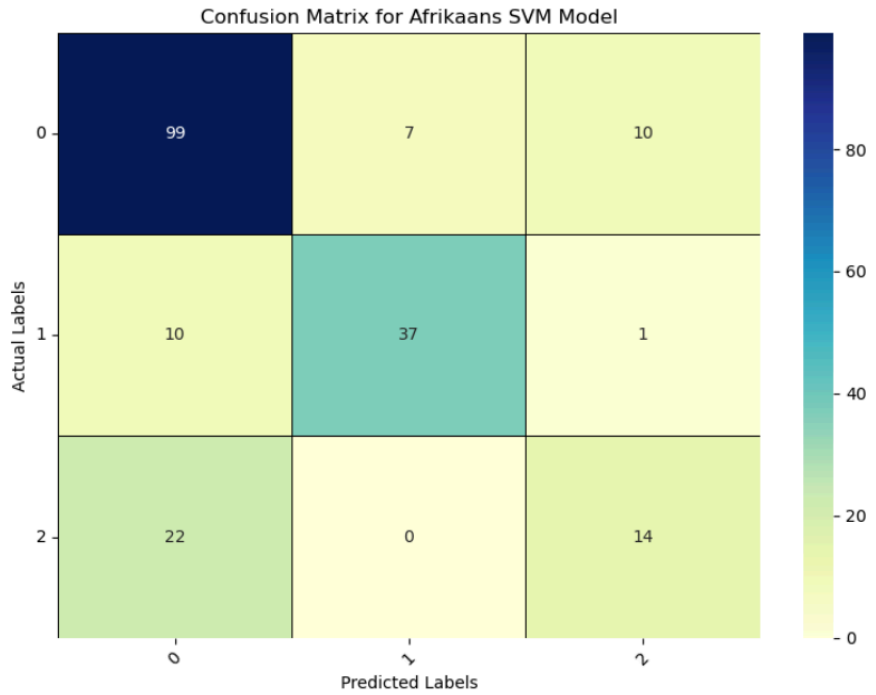


Figure 32. Confusion matrix for afrikaans SVM model

Accuracy: 0.75

	Precision	Recall	F1-score	Support
0	0.76	0.85	0.80	116
1	0.84	0.77	0.80	48
2	0.56	0.39	0.46	36

Table 2. Table of Afrikaans SVM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	nege-en-sestig oomblikke.	0	1
1	vlug vertraag.	0	0
2	drink aborteer.	0	1
3	vlieg verminder.	0	2
4	verkeerd gewapen.	0	0
5	leuen huil.	0	1
6	hartseer siekte.	1	0
7	leuen huil.	0	0
8	leuen huil.	0	0
9	braak smart.	2	0

Figure 33. Afrikaans test sentences with SVM model predictions

Sepedi

For Sepedi, SVM achieved an accuracy of 0.735. This model performed well for class 0, with an F1-score of 0.79, showing strong predictive capabilities for more frequent sentiment classes. The performance dropped significantly for class 2, which had a lower recall, and consequently an F1-score of 0.55. This suggests that the SVM model found it challenging to classify the less frequent classes correctly, resulting in many false negatives.

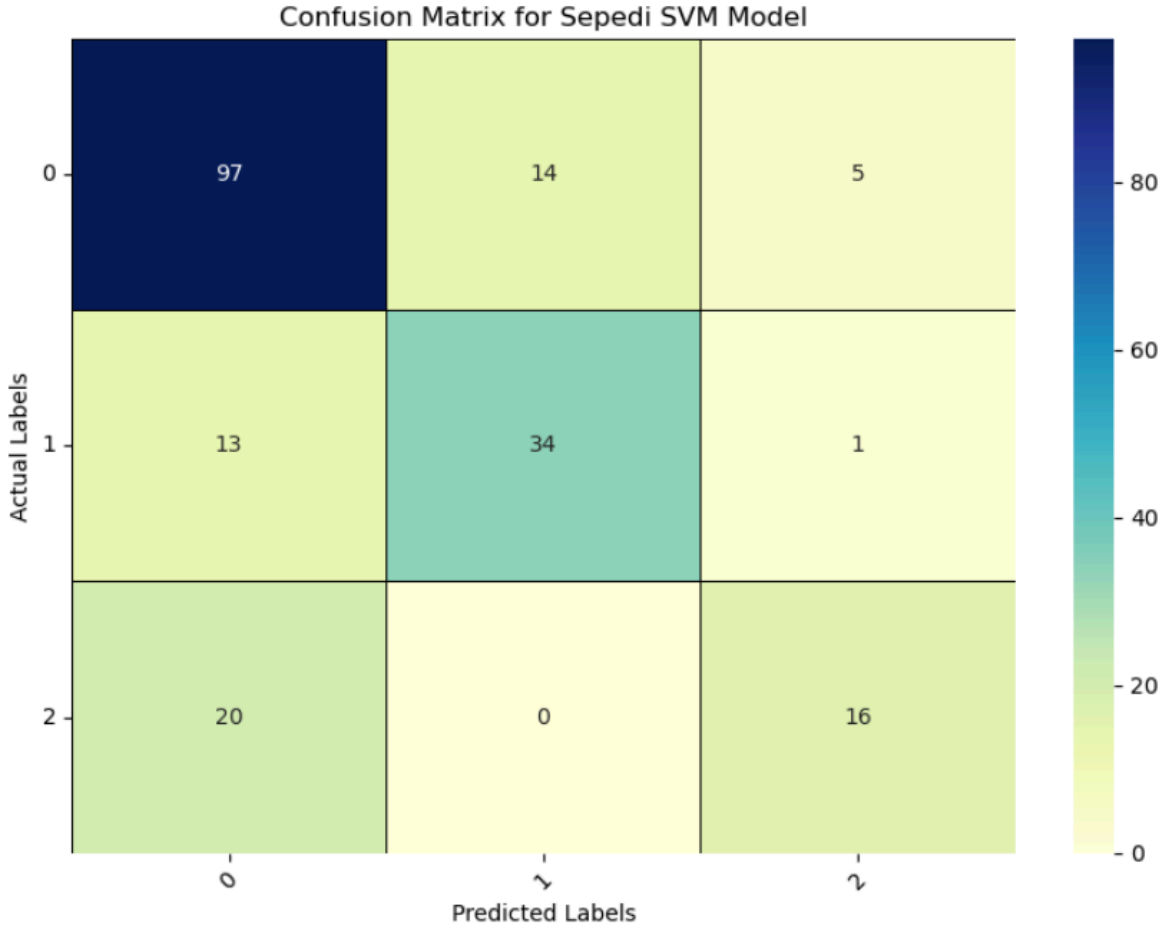


Figure 34. Confusion matrix for Sepedi SVM model

Accuracy: 0.735

	Precision	Recall	F1-score	Support
0	0.75	0.84	0.79	116
1	0.71	0.71	0.71	48
2	0.73	0.44	0.55	36

Table 3. Table of Sepedi SVM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	maiteko a masomehlano-šupa.	0	1
1	go tshwenyega ka lehloyo.	0	0
2	bohloko ba lehlakoreng.	0	0
3	otla lehloyo.	0	2
4	sello go goeletša.	0	0
5	swara gabotse.	0	1
6	manyami a hlasela.	1	0
7	selepe go nyefola.	0	0
8	go palelwa ke go ngangišana.	0	1
9	mmele o opa.	2	0

Figure 35. Sepedi test sentences with SVM model predictions

isiZulu

SVM also achieved an accuracy of 0.735 for Zulu, mirroring its performance on Sepedi. It was particularly good at identifying the majority class (class 0), with an F1-score of 0.79. However, for class 2, recall was lower, leading to poorer performance in identifying less common sentiments accurately. This issue was consistent across all three languages, highlighting the model's difficulty in generalising well for classes with fewer examples.

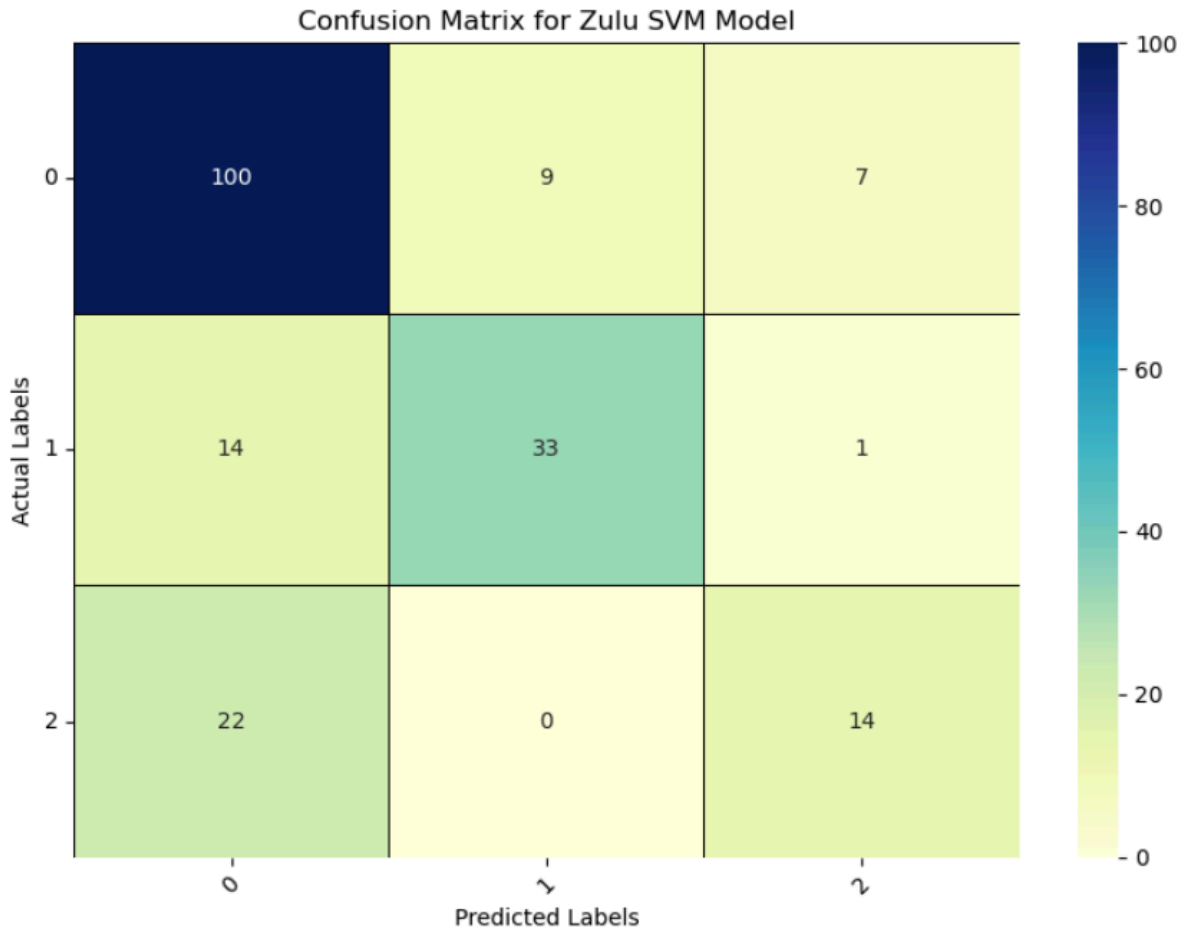


Figure 36. Confusion matrix for isiZulu SVM model

Accuracy: 0.735

	Precision	Recall	F1-score	Support
0	0.74	0.86	0.79	116
1	0.79	0.69	0.73	48
2	0.64	0.39	0.48	36

Table 4. Table of isiZulu SVM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	veza uthando.	0	0
1	nginamanzi	0	0
2	return the book	0	2
3	ijaji mkhuhlane	0	0
4	ngize uhambe	0	0
5	kunjalo	0	1
6	isifebe yenqaba	1	0
7	khala ukuzisola	0	1
8	nkosazana khohlisa	0	0
9	umkhonto imbazo	2	2

Figure 37. isiZulu test sentences with SVM model predictions

English

The SVM model achieved an accuracy of 0.745 for English, indicating strong performance. It was highly effective in distinguishing the Negatif class (class 1), achieving an F1-score of 0.84, and performed well for the Positif class (class 0) with an F1-score of 0.80. However, for the Neutral class (class 2), recall was lower, resulting in an F1-score of 0.39. This reflects the model's struggle with identifying minority sentiments accurately, which is consistent across similar analyses.

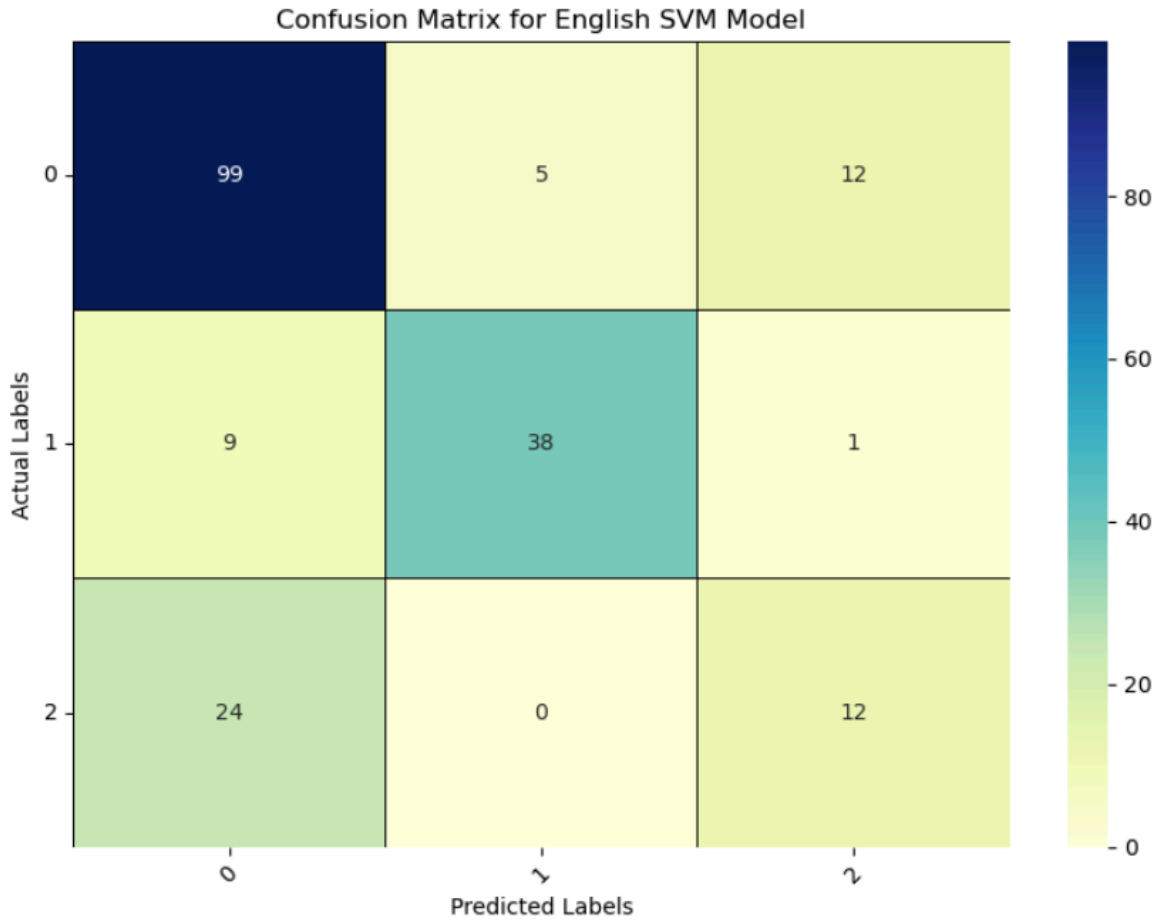


Figure 38. Confusion matrix for English SVM model

Accuracy: 0.745

	Precision	Recall	F1-score	Support
0	0.75	0.85	0.80	116
1	0.88	0.79	0.84	48
2	0.48	0.33	0.39	36

Table 5. Table of English SVM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	armed disease.	0	0
1	sixty-four challenges.	0	0
2	teach peace	0	2
3	buy clothes	0	2
4	bed soft.	0	0
5	stumble lie.	0	0
6	suffer disease.	1	0
7	two options.	0	0
8	drown wet.	0	2
9	lie mourning.	2	0

Figure 39. English test sentences with SVM model predictions

5.1.2.2 CNN

Afrikaans

CNN achieved an accuracy of 0.68, slightly lower than SVM for Afrikaans. It performed well in capturing class 0 and class 1, with F1-scores of 0.74 and 0.73, respectively. The performance for class 2, however, was much weaker, with an F1-score of 0.39. This indicates that CNN struggled with identifying the minority class, especially without sufficient data for class 2.

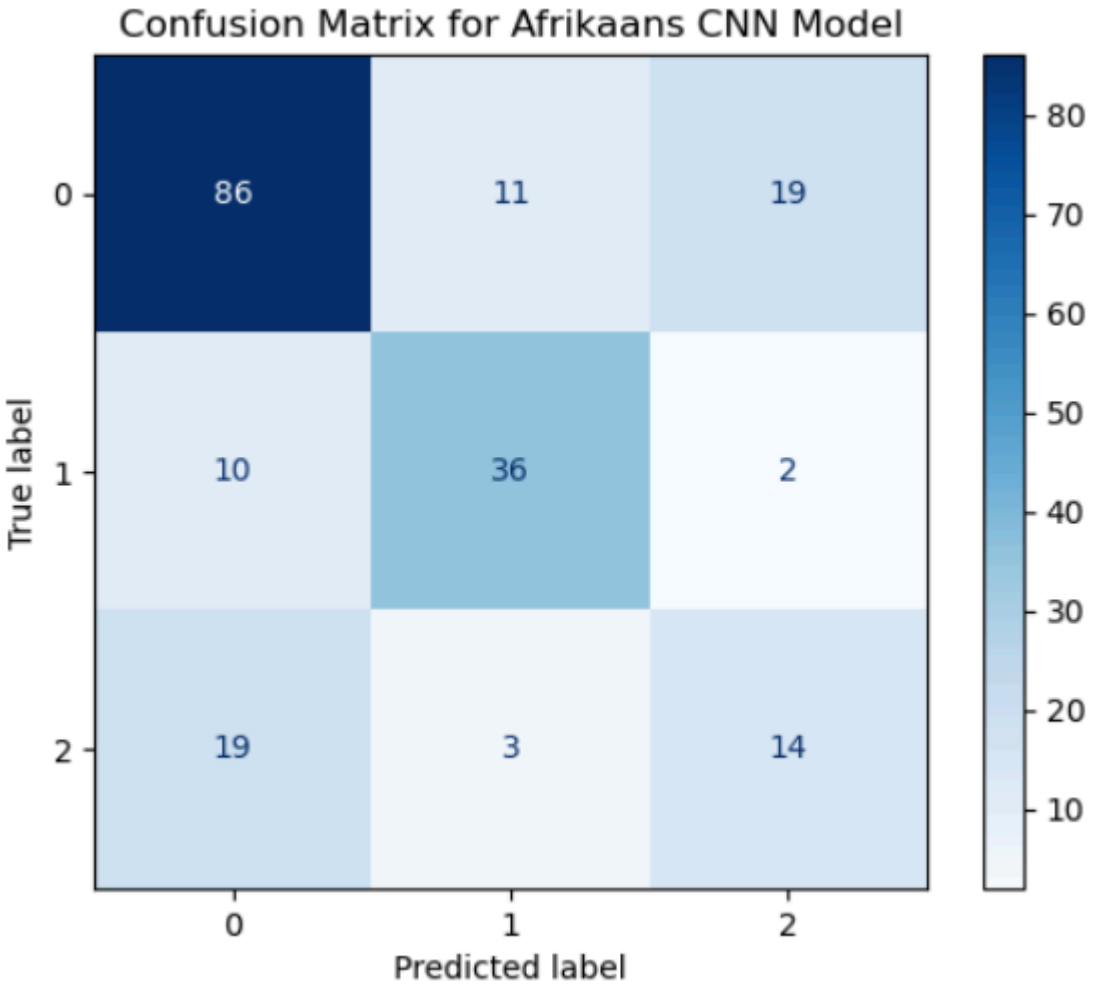


Figure 40. Confusion matrix for Afrikaans CNN model

Accuracy: 0.68

	Precision	Recall	F1-score	Support
0	0.75	0.74	0.74	116
1	0.72	0.75	0.73	48
2	0.40	0.39	0.39	36

Table 6. Table of Afrikaans CNN model metrics

Test Sentences with Predictions:

	Sentence	Predicted Label	Actual
453	vang 'n inbreker	1	0
793	leef sagkens	0	0
209	herhaal dikwels	1	0
309	sewe en sestig probeerslae	2	0
740	bly nederig	0	0
578	sirkel die fout	1	0
895	oorweeg die bruidskat	0	1
545	gaan na die plaas	0	0
436	roep die mense terug	2	0
678	buig stadig	0	2

Figure 41. Afrikaans test sentences with CNN model predictions

Sepedi

The CNN model performed similarly to Afrikaans, with an accuracy of 0.665. It was able to correctly identify most instances of class 0, but the F1-score for class 2 was 0.42. The results suggest that CNN has a harder time capturing the distinctions for the less represented sentiment classes, likely due to a limited number of training examples for these classes. This is consistent with what we typically see in models reliant on hierarchical features and large data.

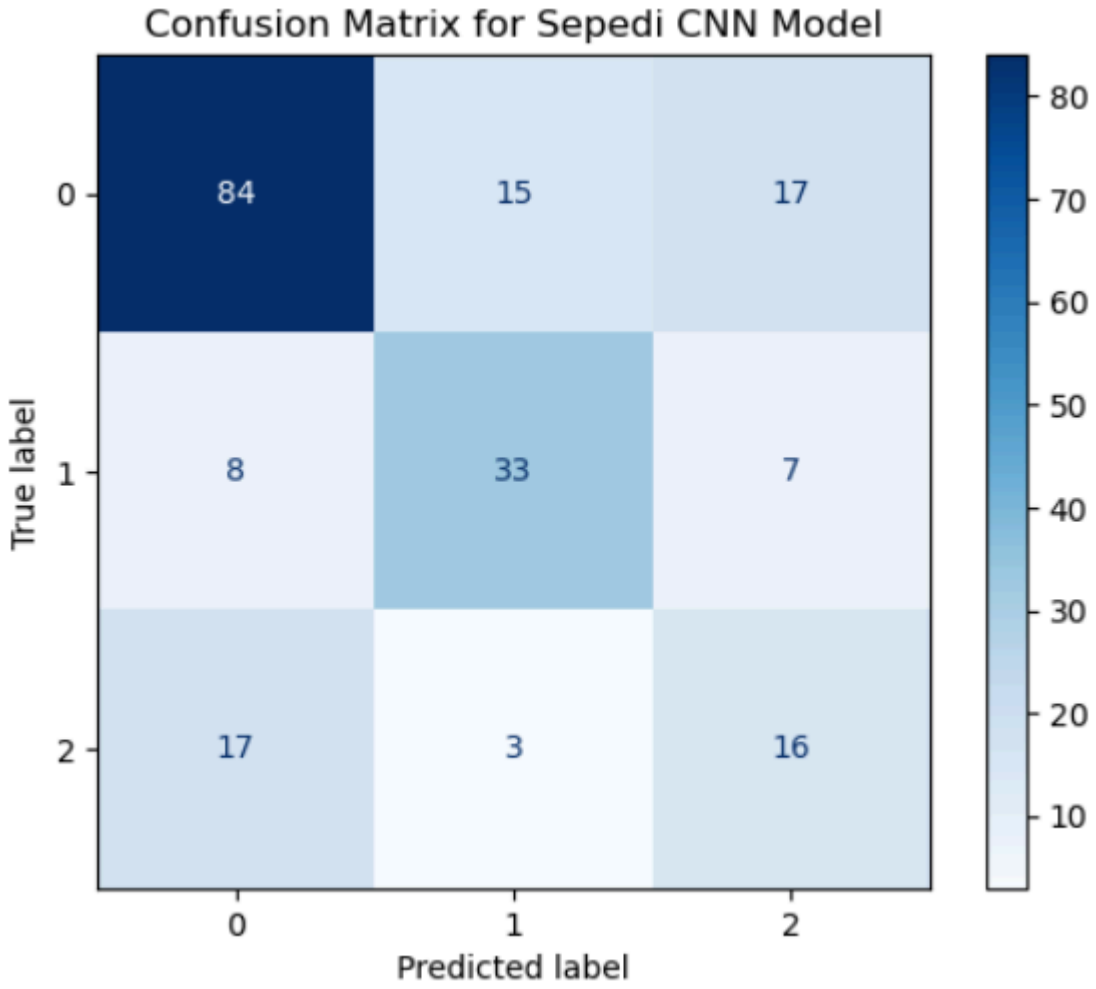


Figure 42. Confusion matrix for Sepedi CNN model

Accuracy: 0.665

	Precision	Recall	F1-score	Support
0	0.77	0.72	0.75	116
1	0.65	0.69	0.67	48
2	0.40	0.44	0.42	36

Table 7. Table of Sepedi CNN model metrics

Test Sentences with Predictions:

	Sentence	Predicted Label	Actual Label
453	hwetša lehodu	0	0
793	phela ka boleta	0	0
209	pheta gantši	0	0
309	diteko tše masometshela šupa	2	0
740	dula o ikokobeditše	0	0
578	sedika phošo	1	0
895	ela hloko magadi	0	1
545	eya tšhemo	0	0
436	bitša batho morago	1	0
678	kobega gannyane	1	2

Figure 43. Sepedi test sentences with CNN model predictions

isiZulu

With an accuracy of 0.67, CNN showed similar performance to Sepedi and Afrikaans. Class 2's F1-score was 0.44, indicating that the model was not able to effectively learn the features distinguishing the minority class. The model performed better on the more frequent classes. For Sepedi, the CNN was more dependent on class frequency, indicating that more data for the minority classes could improve its effectiveness.

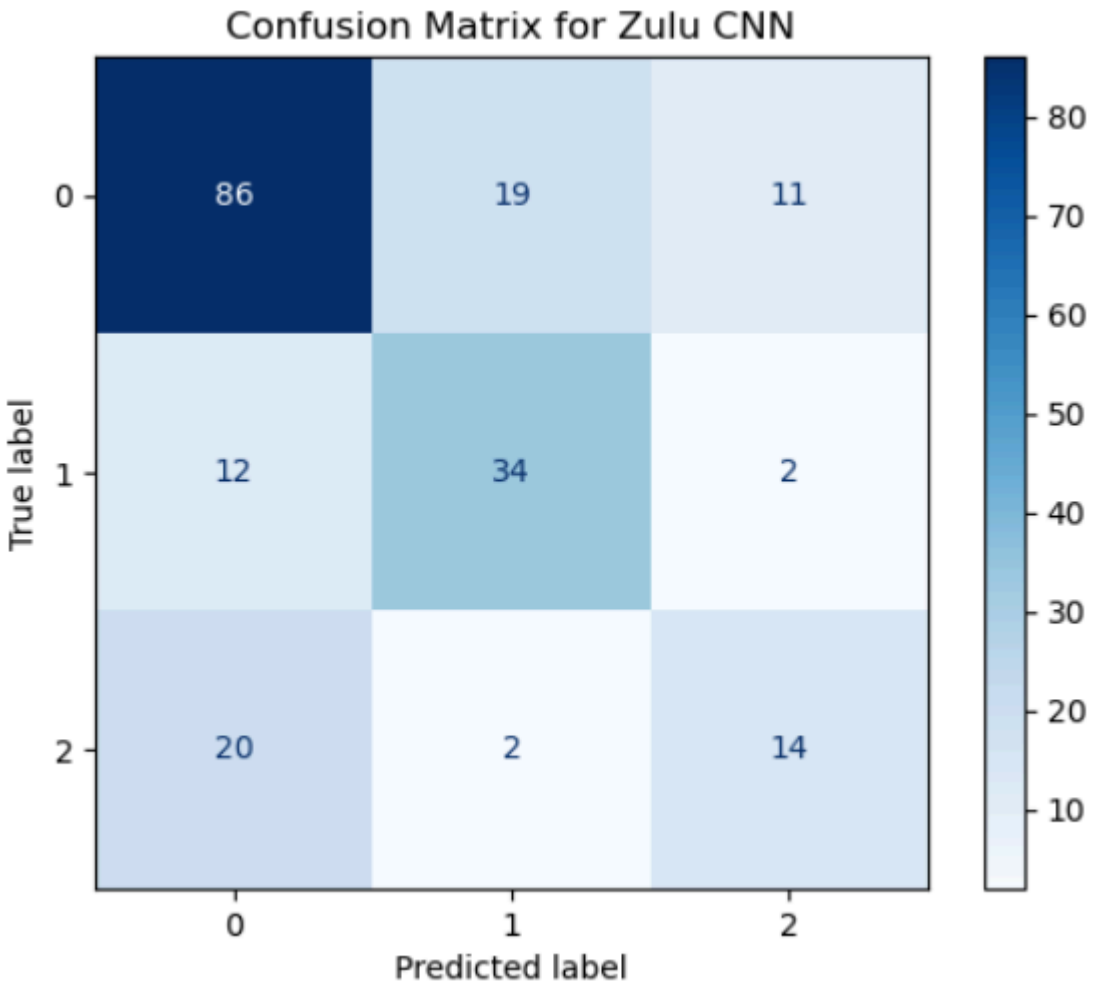


Figure 44. Confusion matrix for isiZulu CNN model

Accuracy: 0.67

	Precision	Recall	F1-score	Support
0	0.73	0.74	0.74	116
1	0.62	0.71	0.66	48
2	0.52	0.39	0.44	36

Table 8. Table of isiZulu CNN model metrics

```
Test Sentences with Predictions:

                                     Sentence  Predicted Label  \
453                            thola umgqekezi                1
793                           phila ngobumnene                0
209                               phinda often                2
309  ama try angamashumi ayisithupha nesikhombisa                0
740                           hlala uthobekile                0
578                            zungeza iphutha                1
895                           cabanga ngedowry                0
545                         hamba uye epulazini                1
436                          biza abantu babuye                1
678                                goba kancane                2

     Actual Label
453             0
793             0
209             0
309             0
740             0
578             0
895             1
545             0
436             0
678             2
```

Figure 45. isiZulu test sentences with CNN model predictions

English

The CNN model achieved an accuracy of 0.68 for English, showing weaker performance compared to SVM. It performed well in predicting the Neutral class (class 1) with an F1-score of 0.80, and Negatif (class 1) with an F1-score of 0.73. However, it struggled significantly with the Positif class (class 0), achieving an F1-score of 0.44, indicating difficulties in capturing the features distinguishing positive sentiments effectively.

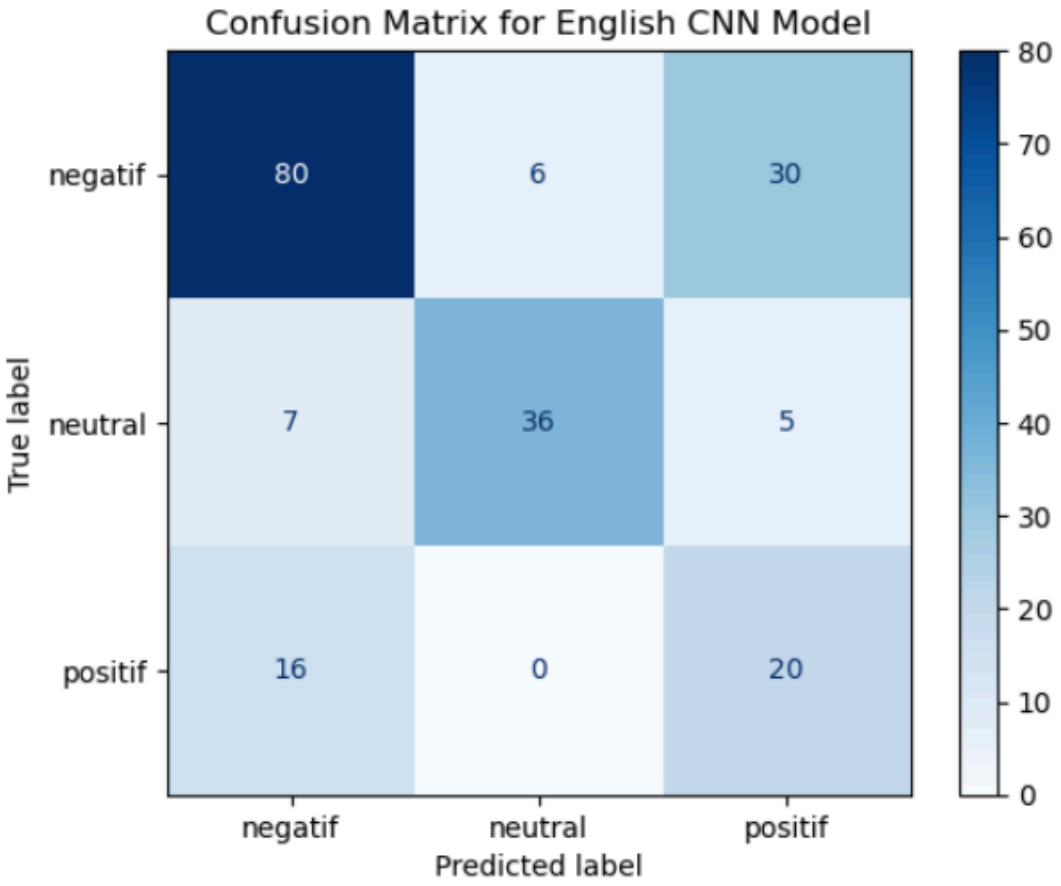


Figure 46. Confusion matrix for English CNN model

Accuracy: 0.68

	Precision	Recall	F1-score	Support
Negatif (1)	0.78	0.69	0.73	116
Neutral (1)	0.86	0.75	0.80	48
Positif (0)	0.36	0.56	0.44	36

Table 9. Table of English CNN model metrics

Test Sentences with Predictions:

	Sentence	Predicted Label	Actual
453	find a burgular	0	0
793	live gently	0	0
209	repeat often	2	0
309	sixty seven tries	2	0
740	stay humble	0	0
578	circle the mistake	1	0
895	consider the dowry	0	1
545	go to the farm	0	0
436	call back the people	2	0
678	bend slowly	2	2

Figure 47. English test sentences with CNN model predictions

5.1.2.3 Naïve Bayes

Afrikaans

Naïve Bayes achieved an accuracy of 0.58, which was the lowest among the three models. The F1-score for the 'Negative' class was 0.69, showing that Naïve Bayes did reasonably well at identifying negative sentiment. However, it performed poorly for the 'Neutral' class, with an F1-score of 0.50, and had difficulty correctly classifying 'Positive' sentiments as well.

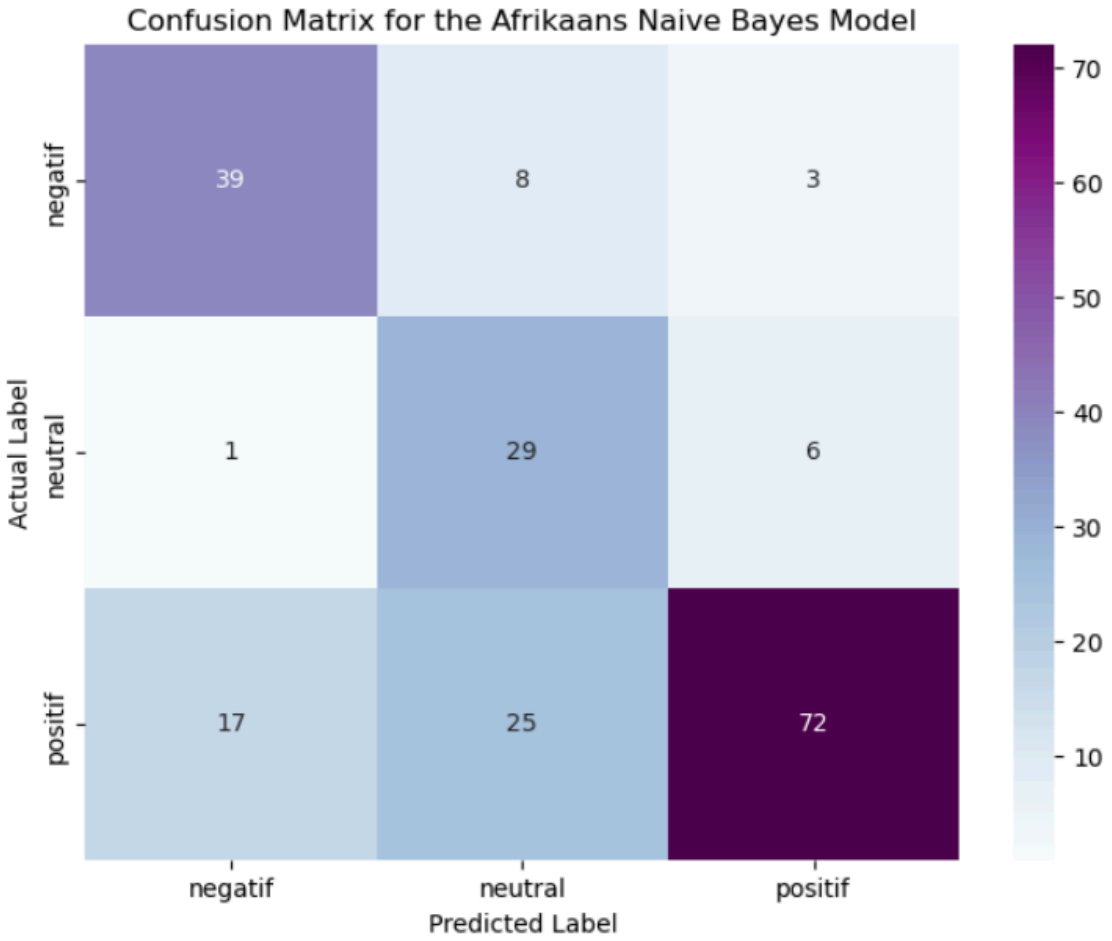


Figure 48. Confusion matrix for the Afrikaans Naive Bayes model

Accuracy: 0.7

	Precision	Recall	F1-score	Support
Negatif (1)	0.68	0..78	0.73	50
Neutral (2)	0.47	0.81	0.59	36
Positif (0)	0.89	0.63	0.74	114

Table 10. Table of Afrikaans Naive Bayes model metric

Test Sentences with Predictions:

	Sentence	Prediction	Actual
453	vang 'n inbreker.	negatif	positif
793	leef sagkens.	positif	positif
209	herhaal dikwels.	neutral	positif
309	sewe-en-sestig probeerslae.	neutral	positif
740	bly nederig.	positif	positif
578	sirkel die fout.	negatif	positif
895	oorweeg die bruidskat.	positif	negatif
545	gaan na die plaas.	negatif	positif
436	roep die mense terug.	positif	positif
678	buig stadig.	neutral	neutral

Figure 49. Afrikaans test sentences with Naive Bayes model predictions

Sepedi

Like its performance on Afrikaans, the Naïve Bayes model had a difficult time trying to achieve higher accuracy for Sepedi and was the least effective among the models tested. It had high recall for neutral sentiments, suggesting it correctly labelled most neutral instances but at the cost of low precision, which means there were many false positives. The model's simplicity meant that it lacked the flexibility to adapt to more complex relationships between words and thus failed to capture some of the nuances of Sepedi sentiment.

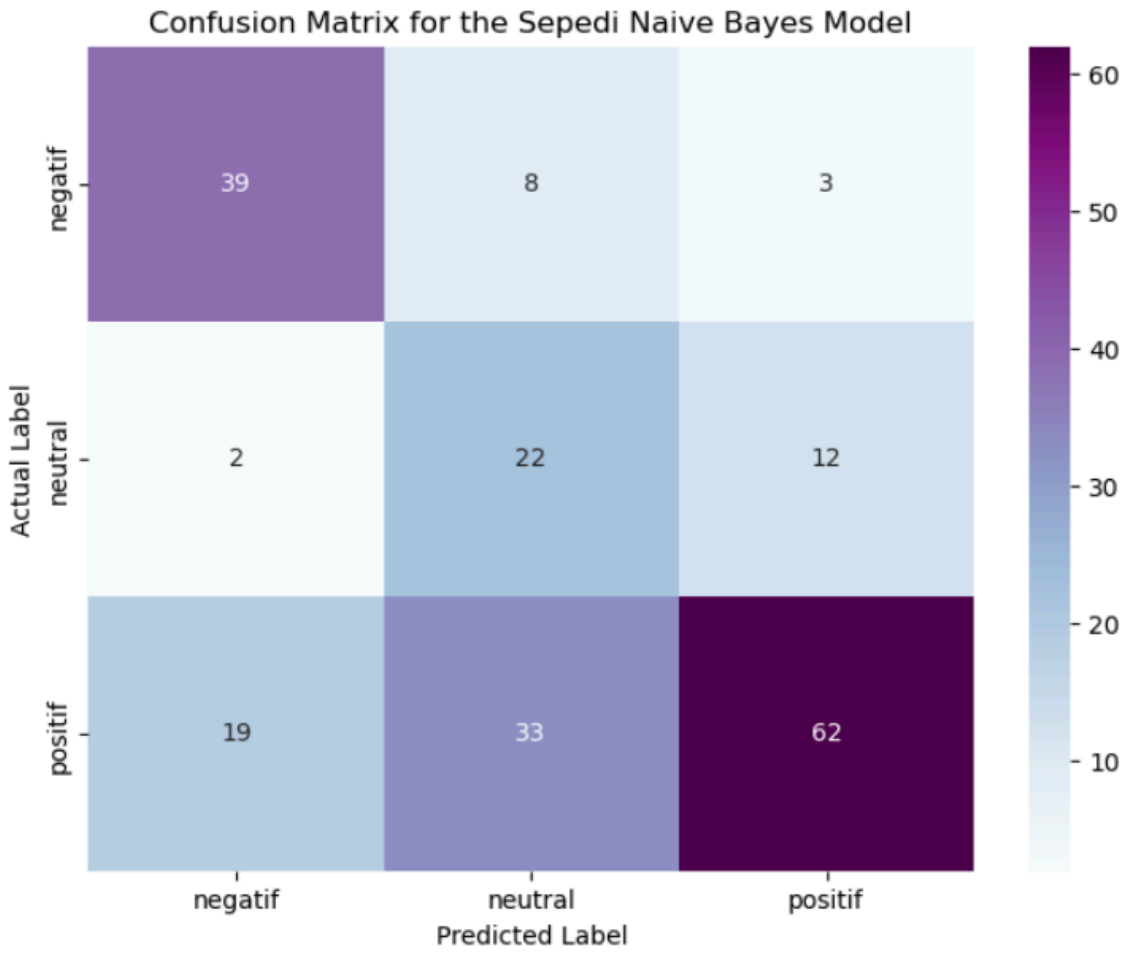


Figure 50. Confusion matrix for the Sepedi Naive Bayes model

Accuracy: 0.615

	Precision	Recall	F1-score	Support
Negatif (1)	0.65	0.78	0.71	50
Neutral (2)	0.35	0.61	0.44	36
Positif (0)	0.81	0.54	0.65	114

Table 11. Table of Sepedi Naive Bayes model metrics

Test Sentences with Predictions:

	Sentence	Prediction	Actual
453	hwetša lehodu.	negatif	positif
793	phela ka boleta.	positif	positif
209	pheta gantši.	neutral	positif
309	diteko tše masometshela-šupa.	neutral	positif
740	dula o ikokobeditše.	neutral	positif
578	sedika phošo.	negatif	positif
895	ela hloko magadi.	positif	negatif
545	eya tšhemo.	negatif	positif
436	bitša batho morago.	positif	positif
678	kobega gannyane.	neutral	neutral

Figure 51. Sepedi test sentences with Naive Bayes model predictions

isiZulu

The Naïve Bayes model's accuracy for Zulu was like the other languages, remaining around 0.58. The recall for classifying neutral sentences was high, but this was paired with a lower precision, indicating that the model tended to overgeneralize neutral classifications. Naïve Bayes was particularly limited in capturing complex relationships and sentiments in Zulu, suggesting that more sophisticated models or additional features are needed for this language.

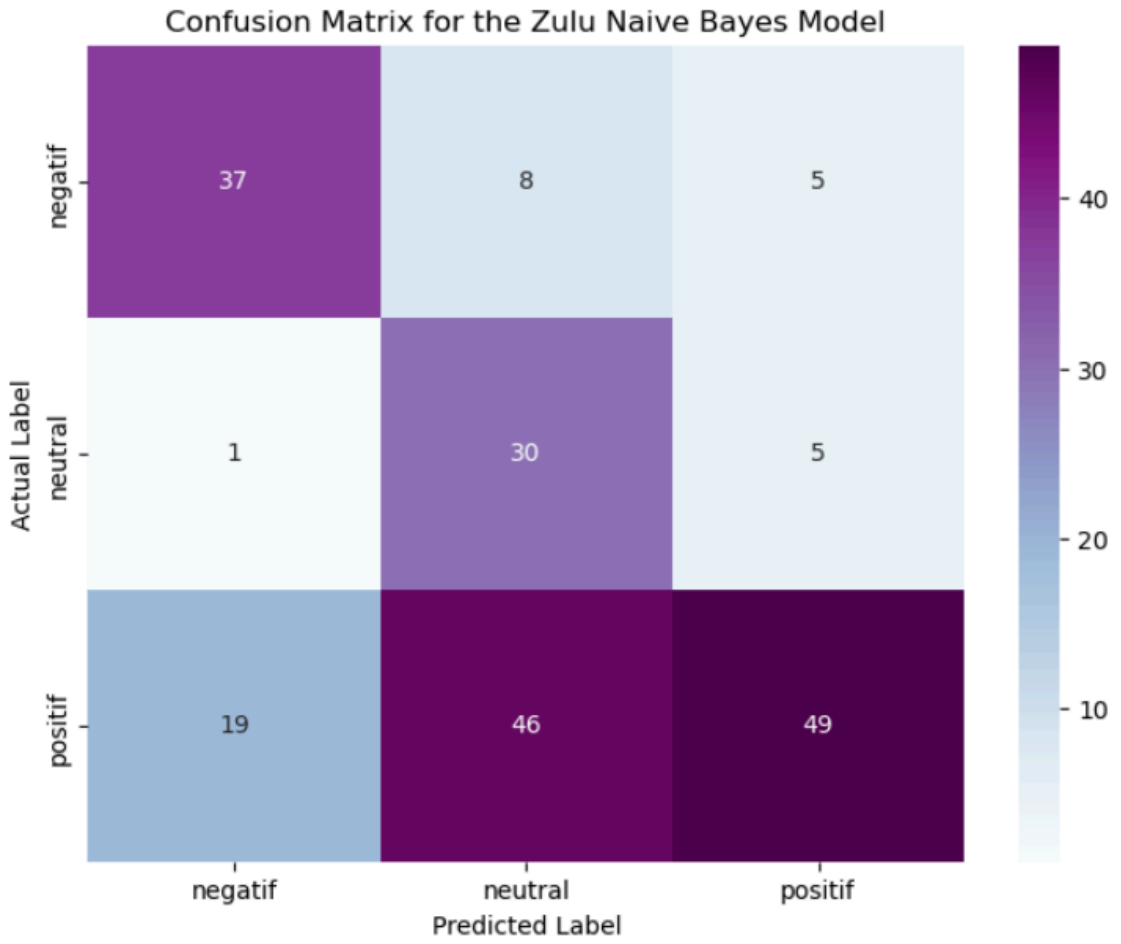


Figure 52. Confusion matrix for isiZulu Naive Bayes model

Accuracy: 0.58

	Precision	Recall	F1-score	Support
Negatif (1)	0.65	0.74	0.69	50
Neutral (2)	0.36	0.83	0.50	36
Positif (0)	0.83	0.43	0.57	114

Table 12. Table of isiZulu Naive Bayes model metrics

Test Sentences with Predictions:

	Sentence	Prediction	Actual
453	thola umgqekezi	negatif	positif
793	phila ngobumnene.	positif	positif
209	phinda often	neutral	positif
309	ama-try angamashumi ayisithupha nesikhombisa	neutral	positif
740	hlala uthobekile	neutral	positif
578	zungeza iphutha	negatif	positif
895	cabanga ngedowry.	positif	negatif
545	hamba uye epulazini	negatif	positif
436	biza abantu babuye	neutral	positif
678	goba kancane	neutral	neutral

Figure 53. isiZulu test sentences with Naive Bayes model predictions

English

The Naïve Bayes model achieved an accuracy of 0.67 for English, which was lower compared to SVM and comparable to CNN. It performed well in identifying Positif (class 0) with a precision of 0.87, although the recall was lower at 0.59, resulting in an F1-score of 0.70. For the Neutral class (class 2), the recall was high (0.75), but precision was lower, resulting in overgeneralization and an F1-score of 0.52. Naïve Bayes struggled to capture more complex relationships within sentiments effectively.

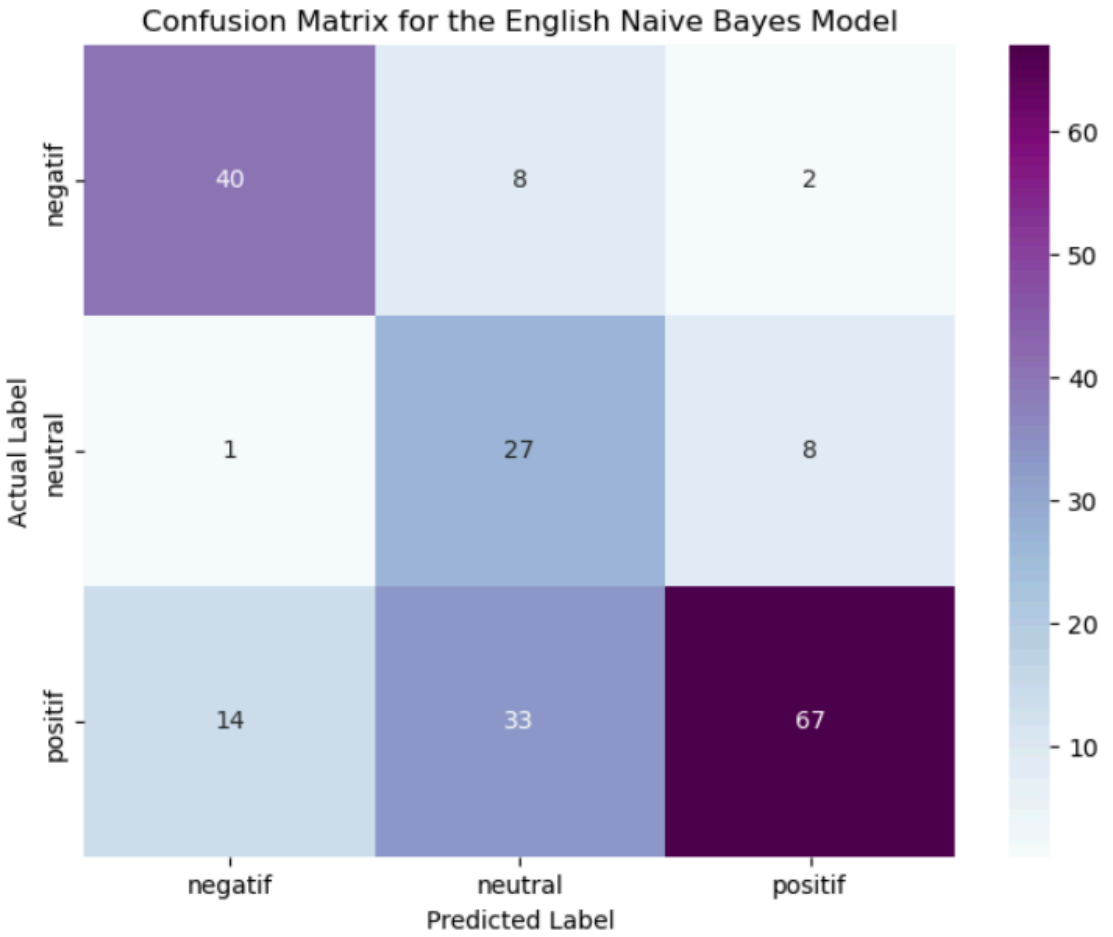


Figure 54. Confusion matrix for English Naive Bayes model

Test Accuracy: 0.67

	Precision	Recall	F1-score	Support
Negatif (1)	0.73	0.80	0.76	50
Neutral (2)	0.40	0.75	0.52	36
Positif (0)	0.87	0.59	0.70	114

Table 13. Table of English Naive Bayes model metrics

Test Sentences with Predictions:

	Sentence	Prediction	Actual
453	find a burgular	positif	positif
793	live gently.	positif	positif
209	repeat often.	neutral	positif
309	sixty-seven tries.	neutral	positif
740	stay humble.	positif	positif
578	circle the mistake	negatif	positif
895	consider the dowry.	positif	negatif
545	go to the farm	neutral	positif
436	call back the people	neutral	positif
678	bend slowly.	neutral	neutral

Figure 55. English test sentences with Naive Bayes model predictions

5.1.2.4. LSTM

Afrikaans

The LSTM model achieved an accuracy of 0.72 for Afrikaans, showing an improvement compared to previous versions, surpassing CNN and approaching the accuracy of SVM. The Positif class (class 0) showed the best performance, with a precision of 0.74, recall of 0.83, and an F1-score of 0.79. The high recall indicates that the model was effective in identifying almost all positive instances, resulting in fewer false negatives. For the Negatif class (class 1), the LSTM achieved a precision of 0.75 and a recall of 0.76, with an F1-score of 0.75. This is a strong performance, suggesting the model was consistent in classifying negative sentiments accurately. The Neutral class (class 2) had the lowest performance, with a precision of 0.52, recall of 0.31, and an F1-score of 0.39. The low recall demonstrates difficulty in identifying neutral sentiments, leading to a significant number of false negatives in this category. Overall, the LSTM model performed well with the Positif and Negatif classes, showing consistent precision and recall. However, it faced challenges in recognizing neutral sentiments effectively.

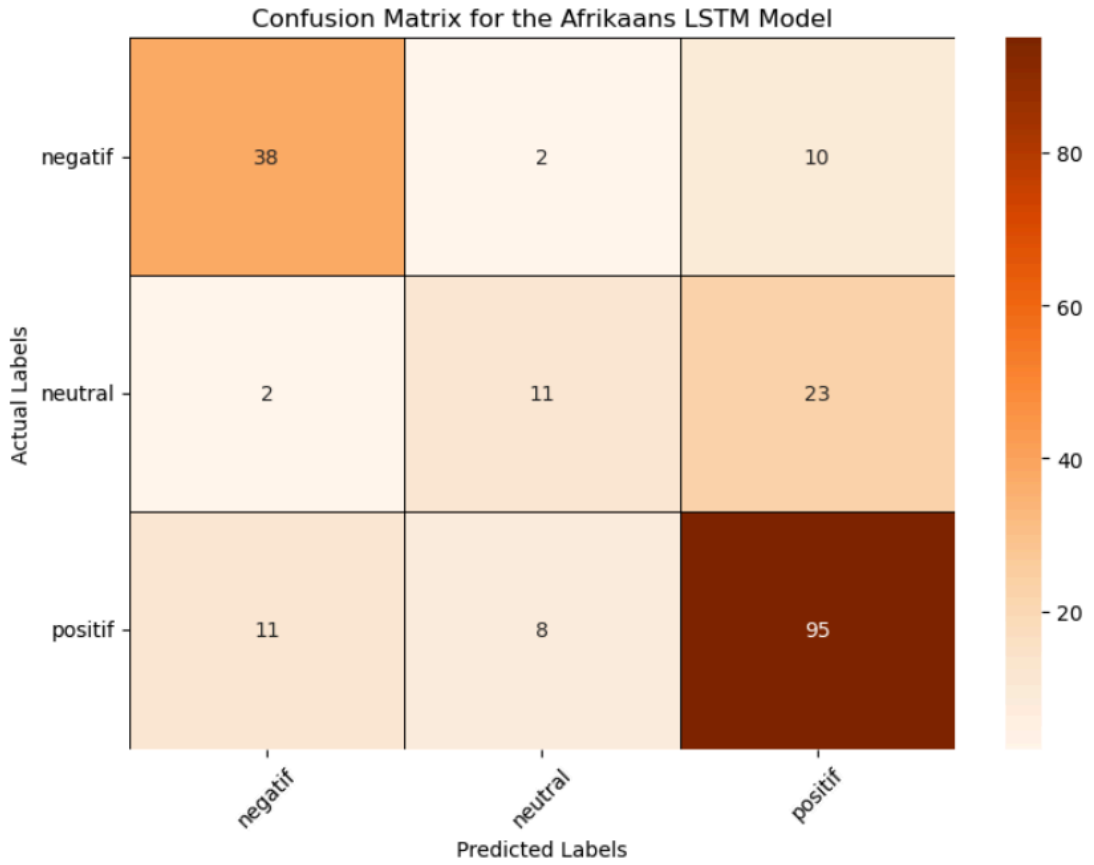


Figure 56. Confusion matrix for Afrikaans LSTM model

Accuracy: 0.72

	Precision	Recall	F1-score	Support
Negatif (1)	0.75	0.76	0.75	50
Neutral (2)	0.52	0.31	0.39	36
Positif (0)	0.74	0.83	0.79	114

Table 14. Table of Afrikaans LSTM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	wanorde jammer.	positif	negatif
1	imbesiliteit haat.	positif	positif
2	skud mis.	positif	negatif
3	haat mis.	positif	neutral
4	jammer tref.	positif	positif
5	tref naby.	positif	negatif
6	mis naby.	negatif	positif
7	mis manipuleer.	positif	positif
8	naby jag.	positif	positif
9	manipuleer kla.	neutral	negatif

Figure 57. Afrikaans test sentences with LSTM model predictions

Sepedi

The LSTM model achieved an accuracy of 0.70 for Sepedi, showing a slight improvement over previous iterations and comparable performance to other models. The Positif class (class 0) had the best performance, with a precision of 0.74, recall of 0.81, and an F1-score of 0.77. This indicates that the model was effective at accurately identifying positive sentiments and minimising false negatives. For the Negatif class (class 1), the model achieved a precision of 0.66 and a recall of 0.70, resulting in an F1-score of 0.68. This is a solid performance, showing that the model managed to correctly classify most negative instances. The Neutral class (class 2) had the weakest performance, with a precision of 0.61, recall of 0.39, and an F1-score of 0.47. The low recall indicates that the model struggled to identify many neutral sentiments, leading to more false negatives in this category. Overall, the LSTM model was most effective at identifying positive sentiments but faced challenges with neutral instances, similar to the performance trend observed in other models across languages.

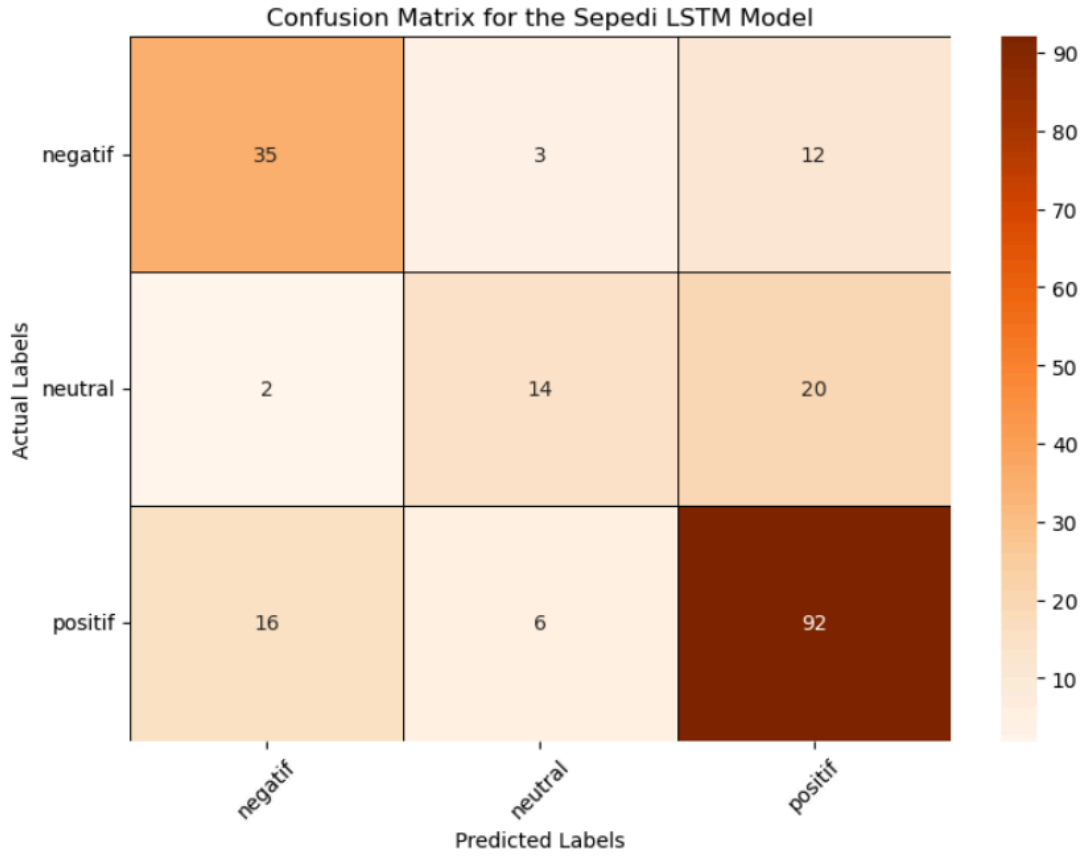


Figure 58. Confusion matrix for Sepedi LSTM model

Accuracy: 0.70

	Precision	Recall	F1-score	Support
Negatif (1)	0.66	0.70	0.68	50
Neutral (2)	0.61	0.39	0.47	36
Positif (0)	0.74	0.81	0.77	114

Table 15. Table of Sepedi LSTM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	tlhakatlhakano kwelobohloko.	positif	positif
1	setlaela lehloyo.	positif	positif
2	šikinya phoša	positif	positif
3	lehloyo phoša	positif	neutral
4	kwelobohloko e otla.	positif	positif
5	otla kgauswi.	positif	negatif
6	phoša kgauswi.	negatif	positif
7	phoša feeketša	positif	positif
8	go tsoma kgauswi.	positif	negatif
9	feeketša sello.	neutral	positif

Figure 59. Sepedi test sentences with LSTM model predictions

isiZulu

The LSTM model achieved an accuracy of 0.68 for Zulu, which is comparable to CNN but slightly lower than SVM. It performed best in predicting the Positif class (class 0), with an F1-score of 0.73, indicating effective identification of positive sentiments. However, for the Neutral class (class 2), the recall was only 0.39, resulting in an F1-score of 0.45, which shows that the model struggled to correctly identify many neutral instances. This challenge was consistent with the LSTM's performance across different classes, demonstrating difficulty with less frequent or more nuanced sentiments.

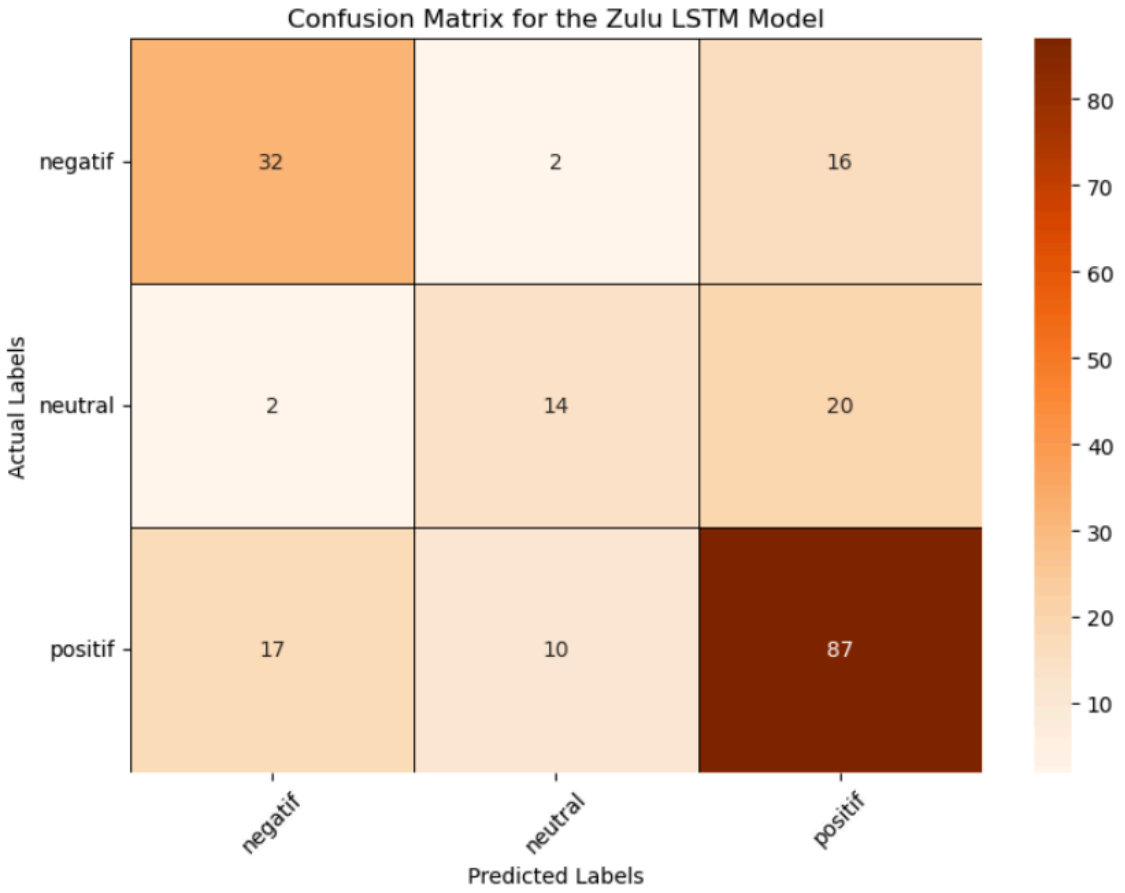


Figure 60. Confusion matrix for isiZulu LSTM model

Accuracy: 0.68

	Precision	Recall	F1-score	Support
Negatif (1)	0.63	0.64	0.63	50
Neutral (2)	0.54	0.39	0.45	36
Positif (0)	0.71	0.76	0.73	114

Table 16. Table of isiZulu LSTM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	inyakanyaka isihawu	positif	negatif
1	ubuwula inzondo	positif	positif
2	shukuza nkosazana	positif	neutral
3	inzondo nkosazana	positif	positif
4	isihawu shaya	positif	positif
5	shaya vala	positif	negatif
6	nkosazana vala	negatif	positif
7	nkosazana khohlisa	positif	negatif
8	vala zingela	positif	negatif
9	khohlisa khala	neutral	positif

Figure 61. isiZulu test sentences with LSTM model predictions

English

The LSTM model achieved an accuracy of 0.715 for English, which was better than CNN and close to SVM. The model performed best for the Positif class (class 0), with an F1-score of 0.79, showing effective identification of positive sentiments. However, it struggled with the Neutral class (class 2), with a recall of 0.22 and an F1-score of 0.29, highlighting difficulties in recognizing neutral sentiments consistently. This suggests the model had issues with less frequent and more nuanced sentiments, similar to the other models' challenges.

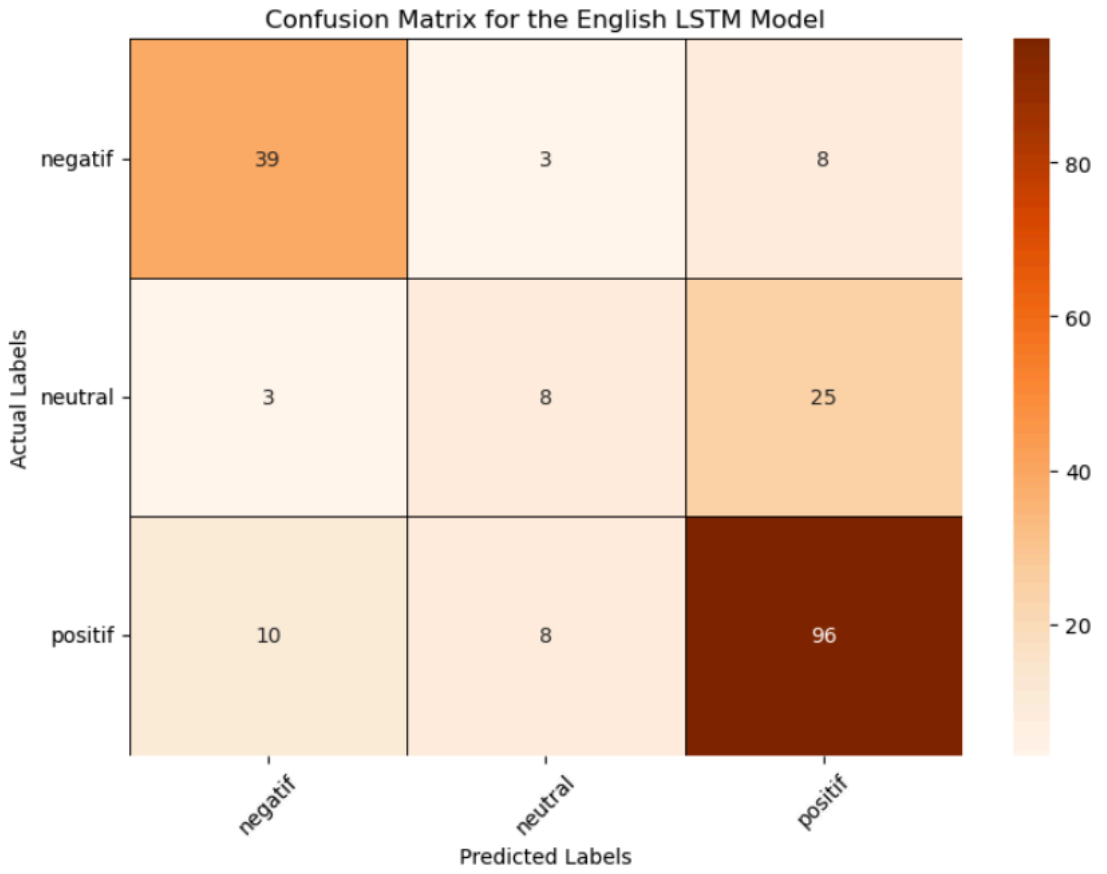


Figure 62. Confusion matrix for English LSTM model

Accuracy: 0.715

	Precision	Recall	F1-score	Support
Negatif (1)	0.75	0.78	0.76	50
Neutral (2)	0.42	0.22	0.29	36
Positif (0)	0.74	0.84	0.79	114

Table 17. Table of English LSTM model metrics

Test Sentences with Predictions:

	Test Sentence	Actual Label	Predicted Label
0	disorder pity.	positif	positif
1	imbecility hate.	positif	positif
2	shake miss.	positif	positif
3	hate miss.	positif	neutral
4	pity hit.	positif	positif
5	hit close.	positif	negatif
6	miss close.	negatif	positif
7	miss manupulate.	positif	positif
8	close hunting.	positif	neutral
9	manupulate lament.	neutral	positif

Figure 63. English test sentences with LSTM model predictions

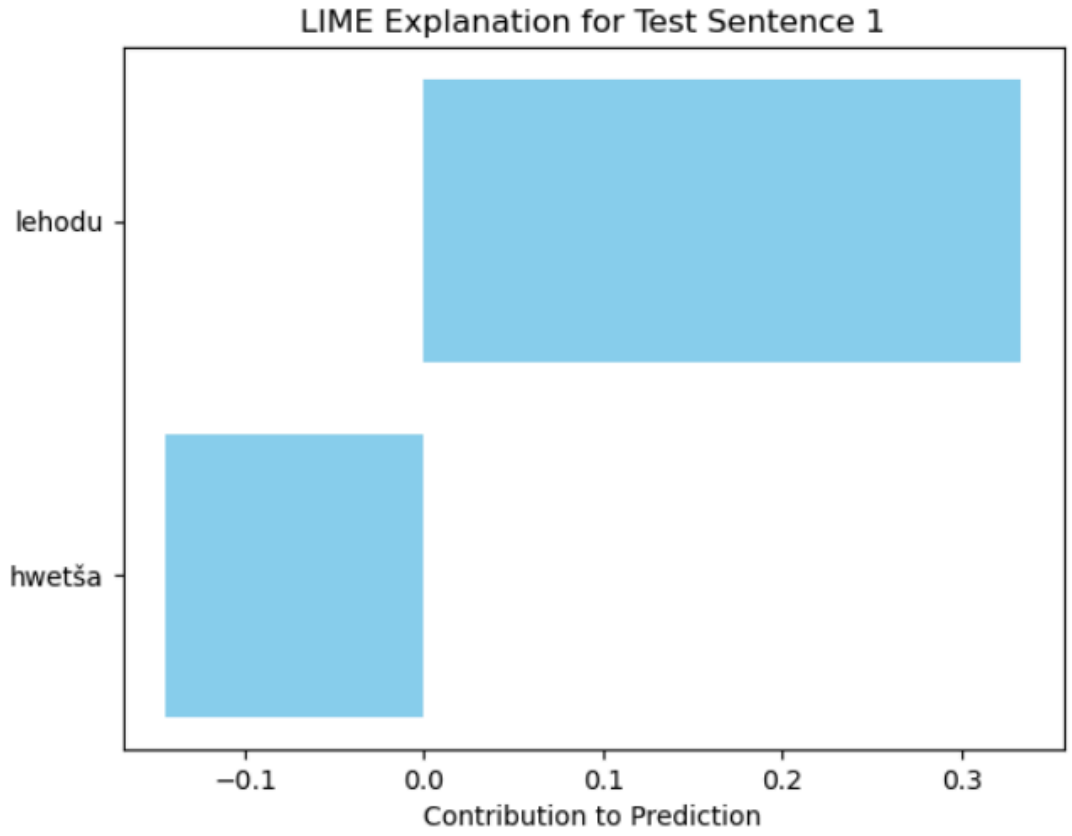
5.1.2.5 Summary of Model Performance Across Languages

Support Vector Machine (SVM) consistently performed best, with accuracies ranging from 0.735 to 0.745 across languages. It did a great job identifying the majority sentiments but had some trouble with the less common, Neutral class. Convolutional Neural Network (CNN) performed moderately well, with accuracies between 0.67 and 0.68. It performed well for more frequent sentiment classes but faced challenges with minority classes. Naïve Bayes was more of a baseline model, with accuracies around 0.58 to 0.67. It had high recall for some classes but overgeneralized, which led to weaker precision and difficulty handling complex sentiment relationships. Long Short-Term Memory (LSTM) performed a bit better than CNN, with accuracies between 0.68 and 0.72. It was effective for Positif sentiments but struggled with Neutral classes, indicating difficulties with subtle and less frequent sentiments. Overall, SVM emerged as the most effective model, while CNN and LSTM showed potential that could be enhanced with more balanced data. Naïve Bayes was helpful as a starting point but lacked the capability to fully capture the nuances in sentiment analysis.

5.2 XAI using LIME

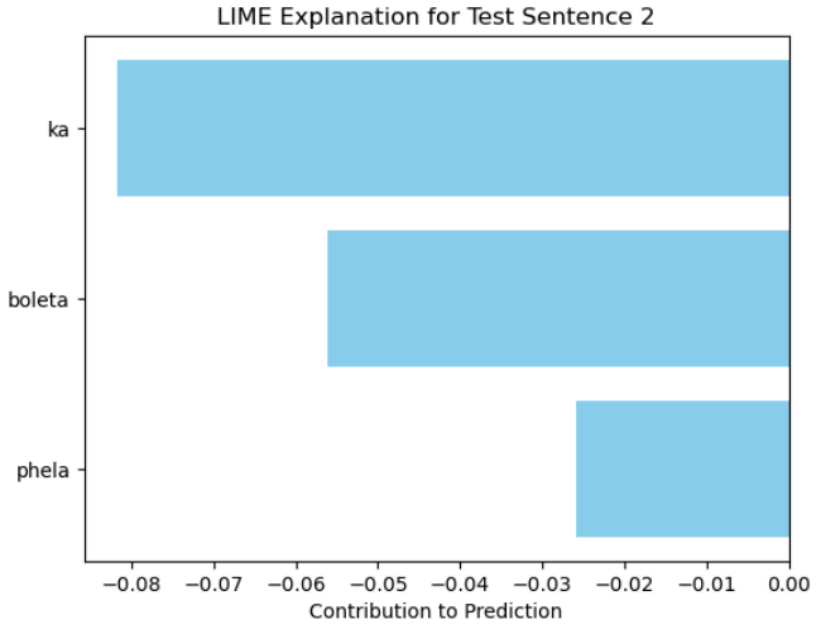
Five test sentences were used for the XAI to indicate how machine learning made its decisions on the classification of the sentences. The best performing model, which was the SVM, was used as the model to explain. Sentences from each language group were used to understand how the predictions were made in each language.

5.2.1 Sepedi Sentences



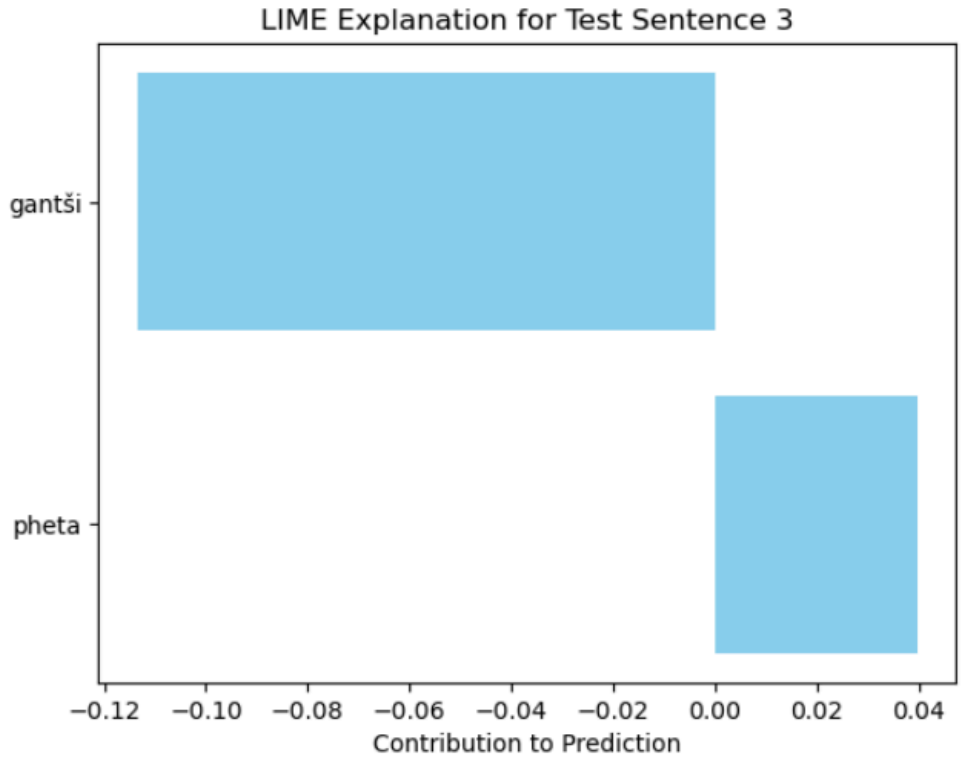

```
Test Sentence 1: hwetša lehodu.
Actual Label: 0
Predicted Label: 1
Explanation: [('lehodu', 0.3329714341444462), ('hwetša', -0.14428264283372508)]
```

Figure 64. Sepedi Sentence 1 Explanation



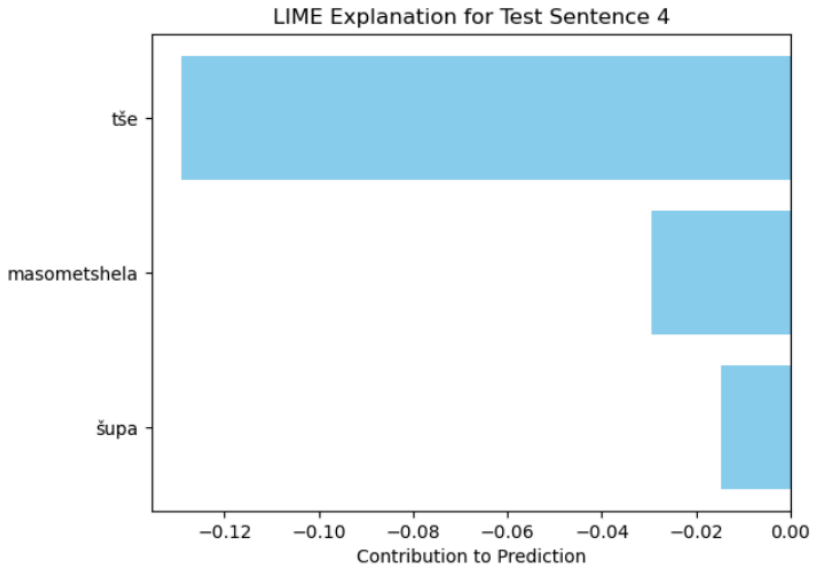

```
Test Sentence 2: phela ka boleta.
Actual Label: 0
Predicted Label: 0
Explanation: [('ka', -0.08169343330901409), ('boleta', -0.0561405891008537), ('phela', -0.025892683685161715)]
```

Figure 65. Sepedi Sentence 2 Explanation



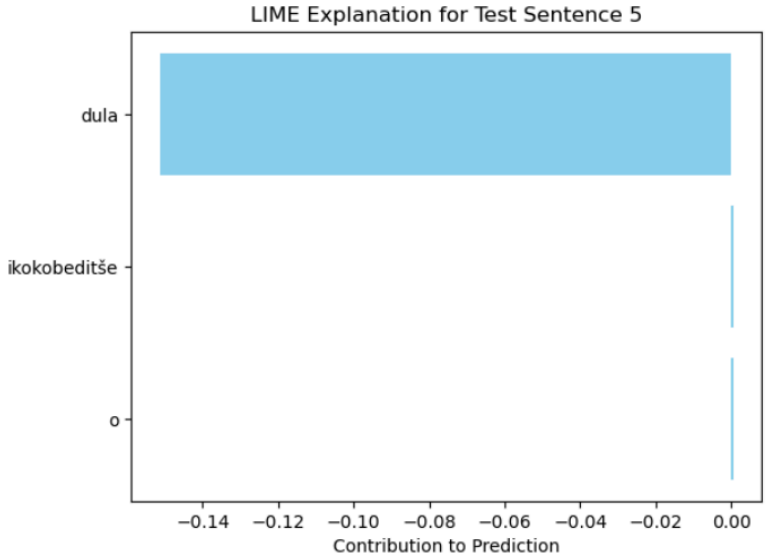

```
Test Sentence 3: pheta gantši.
Actual Label: 0
Predicted Label: 0
Explanation: [('gantši', -0.11338600670107644), ('pheta', 0.039798441917245596)]
```

Figure 66. Sepedi Sentence 3 Explanation




```
Test Sentence 4: diteko tše masometshela-šupa.
Actual Label: 0
Predicted Label: 2
Explanation: [('tše', -0.1290174474391895), ('masometshela', -0.029327593290803377), ('šupa', -0.014695177298295137)]
```

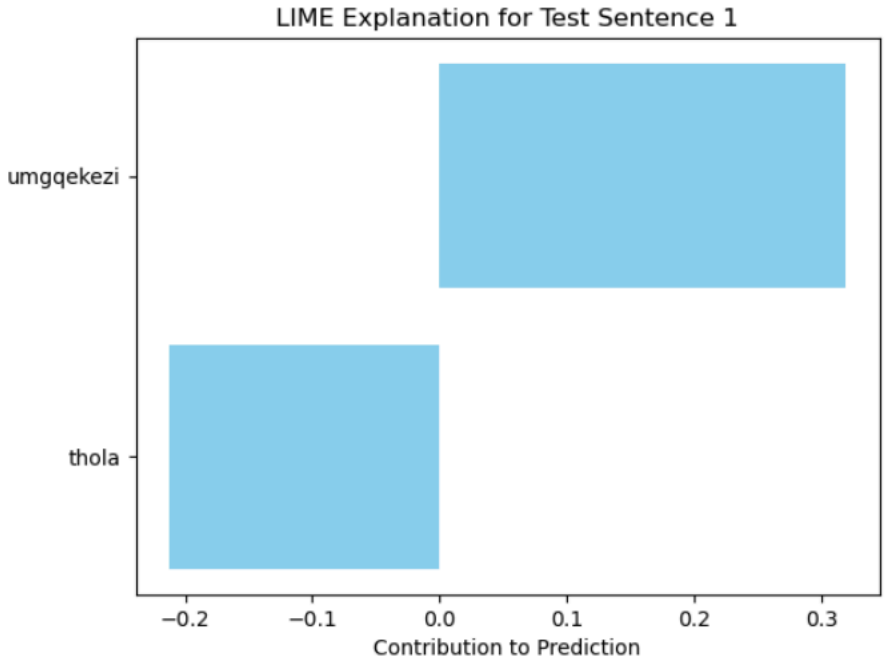
Figure 67. Sepedi Sentence 4 Explanation




```
Test Sentence 5: dula o ikokobeditše.
Actual Label: 0
Predicted Label: 0
Explanation: [('dula', -0.15134349390593088), ('ikokobeditše', 0.0004171047484463626), ('o', 0.0004048664264725715)]
```

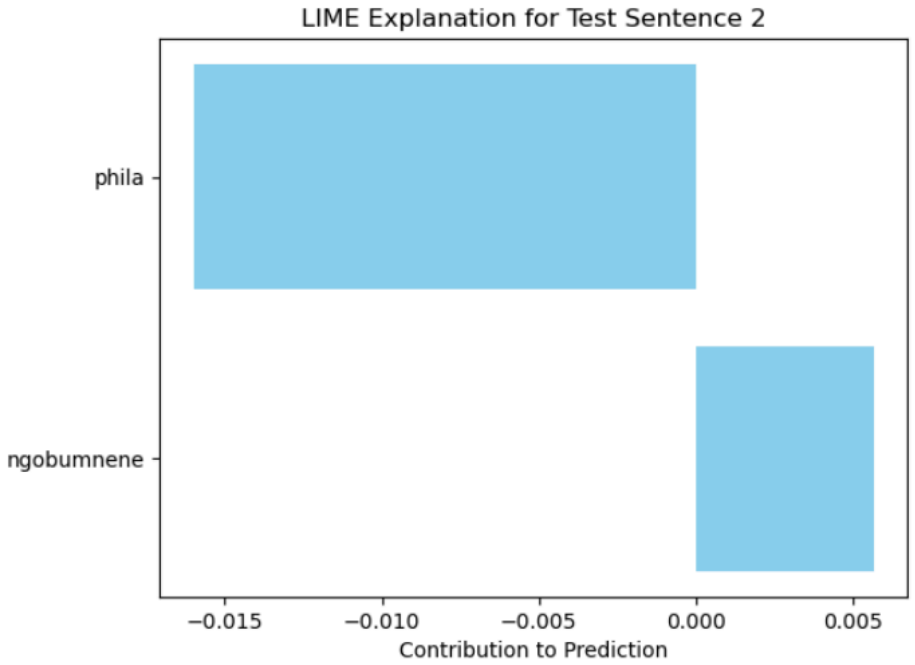
Figure 68. Sepedi Sentence 5 Explanation

5.2.2 Zulu Sentences



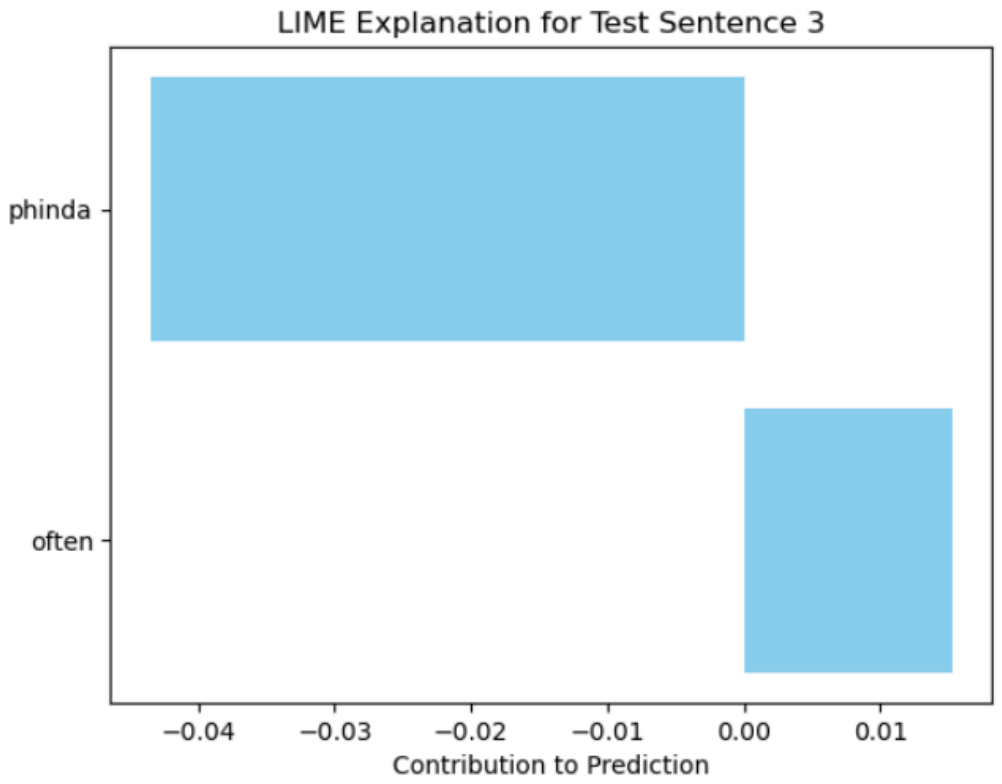

```
Test Sentence 1: thola umgqekezi
Actual Label: 0
Predicted Label: 0
Explanation: [('umgqekezi', 0.31890192391334454), ('thola', -0.21230788523885755)]
```

Figure 69. Zulu Sentence 1 Explanation



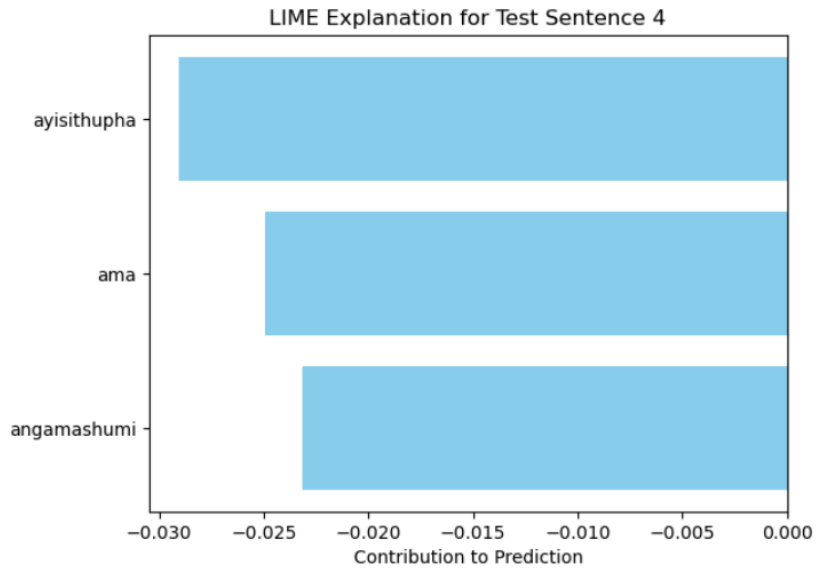

```
Test Sentence 2: phila ngobumnene.
Actual Label: 0
Predicted Label: 0
Explanation: [('phila', -0.01598105155642925), ('ngobumnene', 0.005668052150138012)]
```

Figure 70. Zulu Sentence 2 Explanation



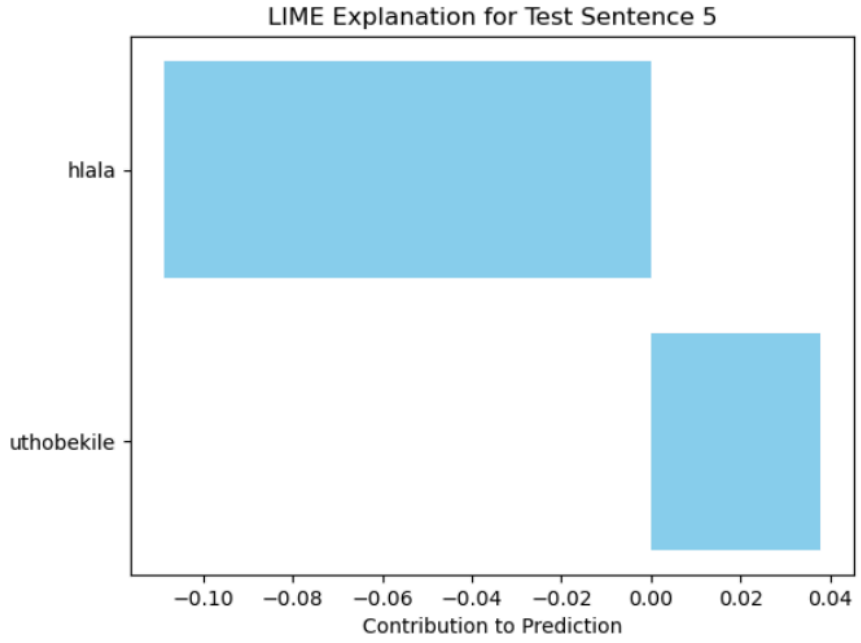

```
Test Sentence 3: phinda often
Actual Label: 0
Predicted Label: 2
Explanation: [('phinda', -0.04349036450936595), ('often', 0.01529787735870092)]
```

Figure 71. Zulu Sentence 3 Explanation




```
Test Sentence 4: ama-try angamashumi ayisithupha nesikhombisa
Actual Label: 0
Predicted Label: 0
Explanation: [('ayisithupha', -0.029035709753493476), ('ama', -0.02492080904393401), ('angamashumi', -0.02318204421314923)]
```

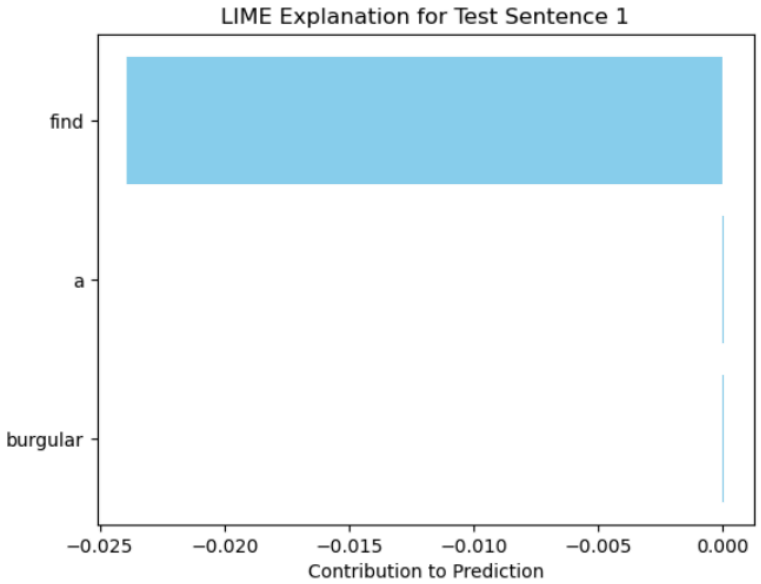
Figure 72. Zulu Sentence 4 Explanation




```
Test Sentence 5: hlala uthobekile
Actual Label: 0
Predicted Label: 0
Explanation: [('hlala', -0.10886437969624418), ('uthobekile', 0.03799854825502721)]
```

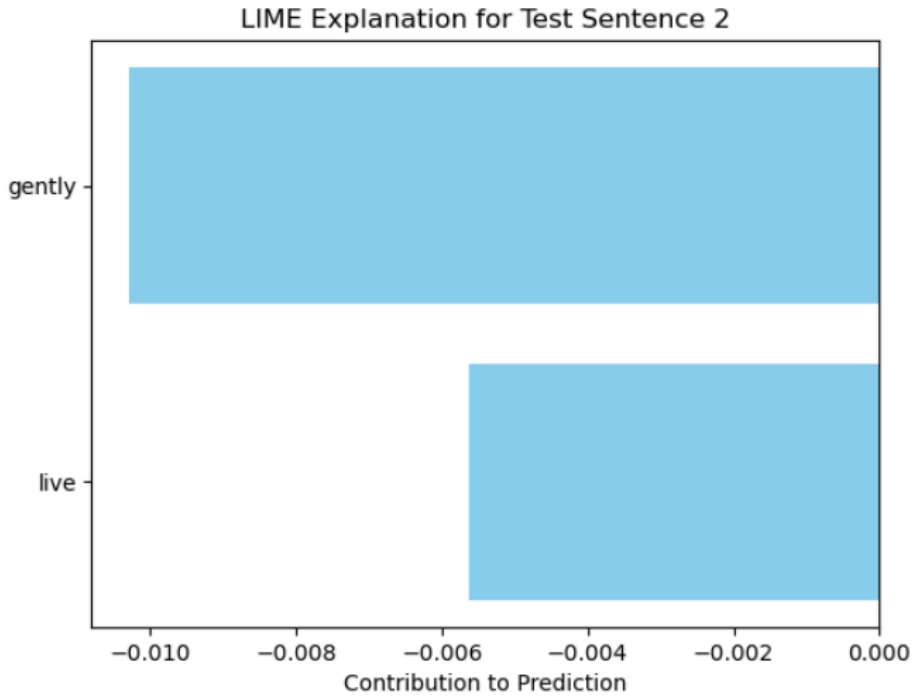
Figure 73. Zulu Sentence 5 Explanation

5.2.3 English Sentences



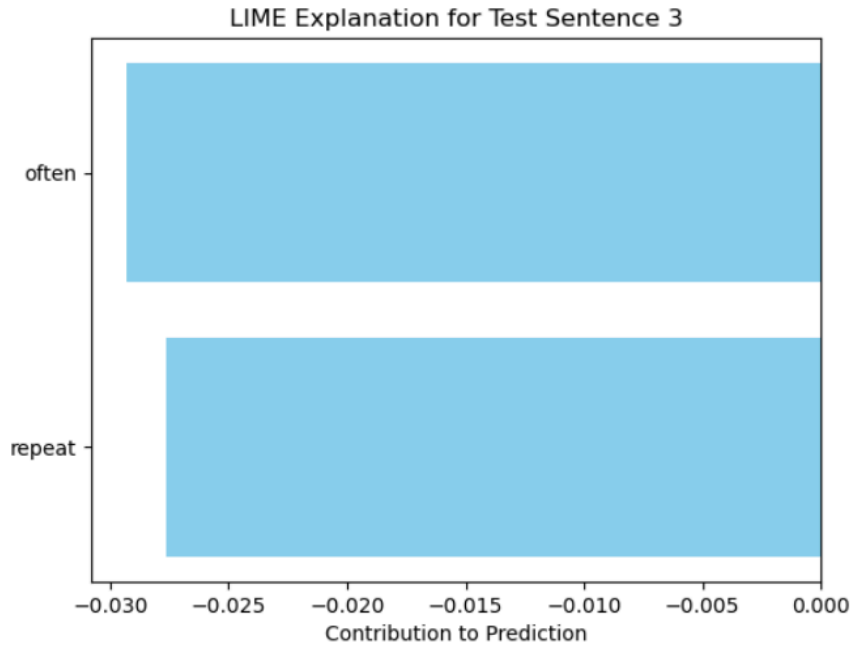

```
Test Sentence 1: find a burgular
Actual Label: 0
Predicted Label: 0
Explanation: [('find', -0.023907466768760283), ('a', 6.353783959975188e-05), ('burgular', 6.253497804991807e-05)]
```

Figure 74. English Sentence 1 Explanation



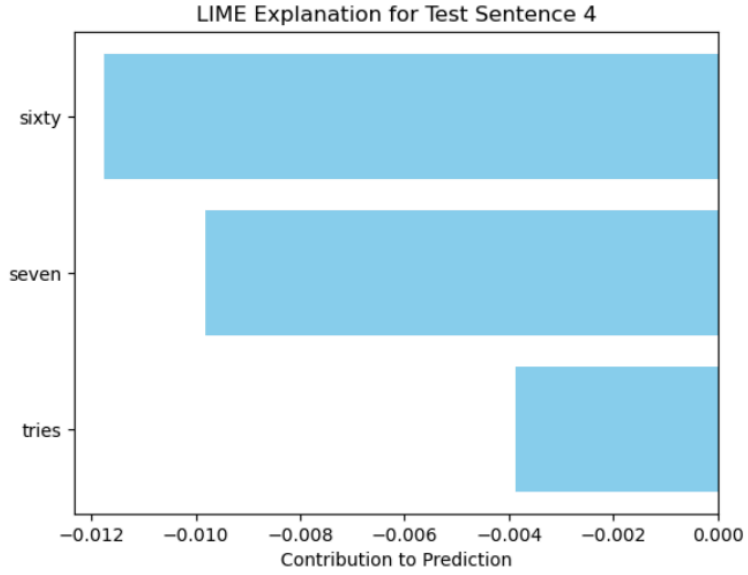

```
Test Sentence 2: live gently.
Actual Label: 0
Predicted Label: 0
Explanation: [('gently', -0.010279892313347756), ('live', -0.005629050374193889)]
```

Figure 75. English Sentence 2 Explanation



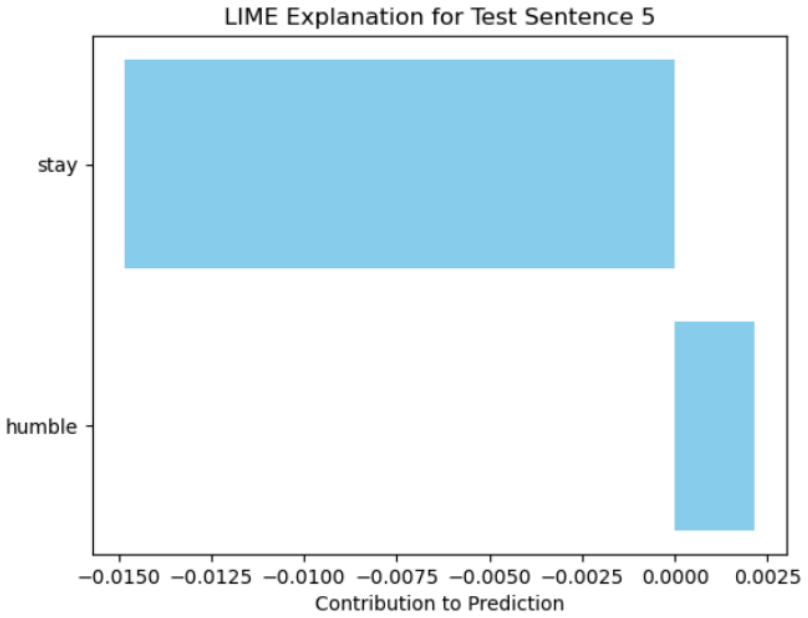

```
Test Sentence 3: repeat often.
Actual Label: 0
Predicted Label: 2
Explanation: [('often', -0.029320652656702625), ('repeat', -0.027665334104498073)]
```

Figure 76. English Sentence 3 Explanation




```
Test Sentence 4: sixty-seven tries.
Actual Label: 0
Predicted Label: 2
Explanation: [('sixty', -0.011742491942222757), ('seven', -0.009818332783676836), ('tries', -0.0038764145723272953)]
```

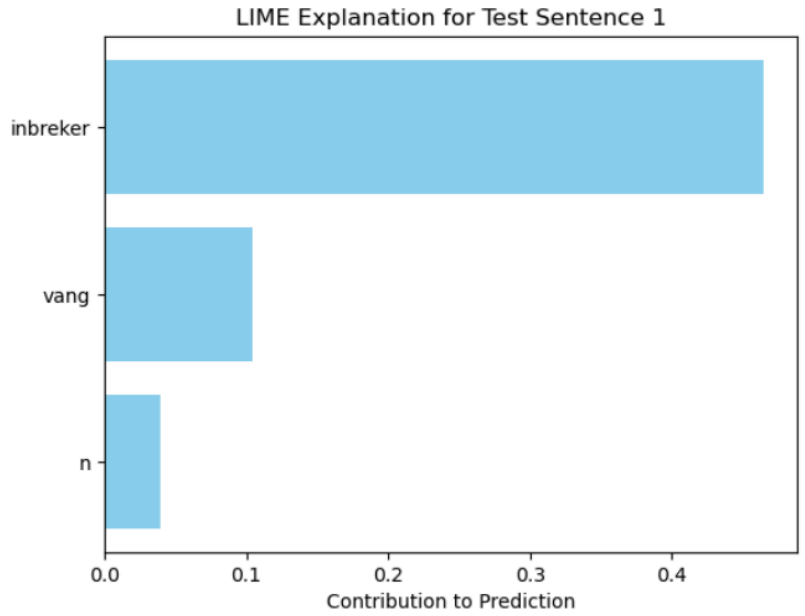
Figure 77. English Sentence 4 Explanation




```
Test Sentence 5: stay humble.
Actual Label: 0
Predicted Label: 0
Explanation: [('stay', -0.014853867922458127), ('humble', 0.0021672063816429348)]
```

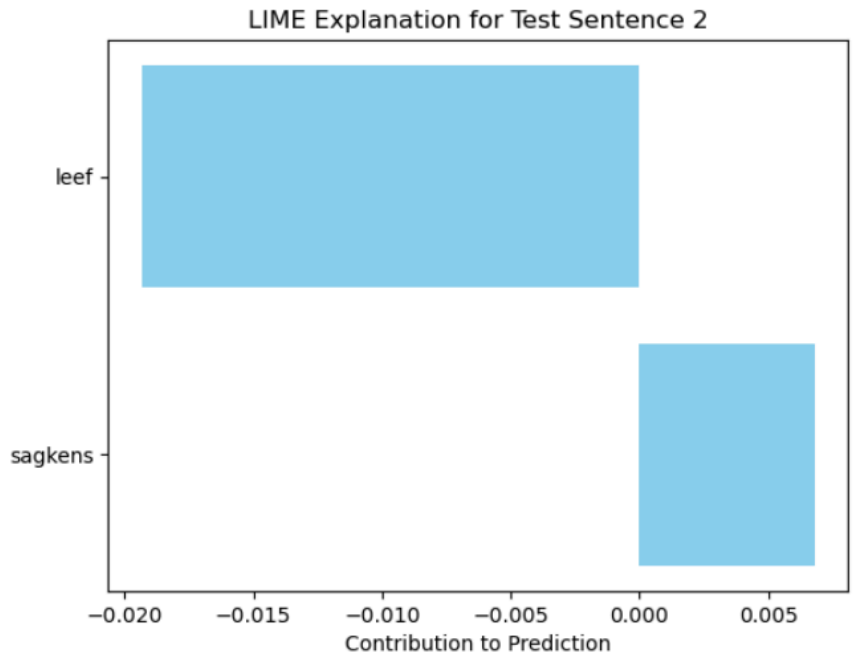
Figure 78. English Sentence 5 Explanation

5.2.4 Afrikaans Sentences



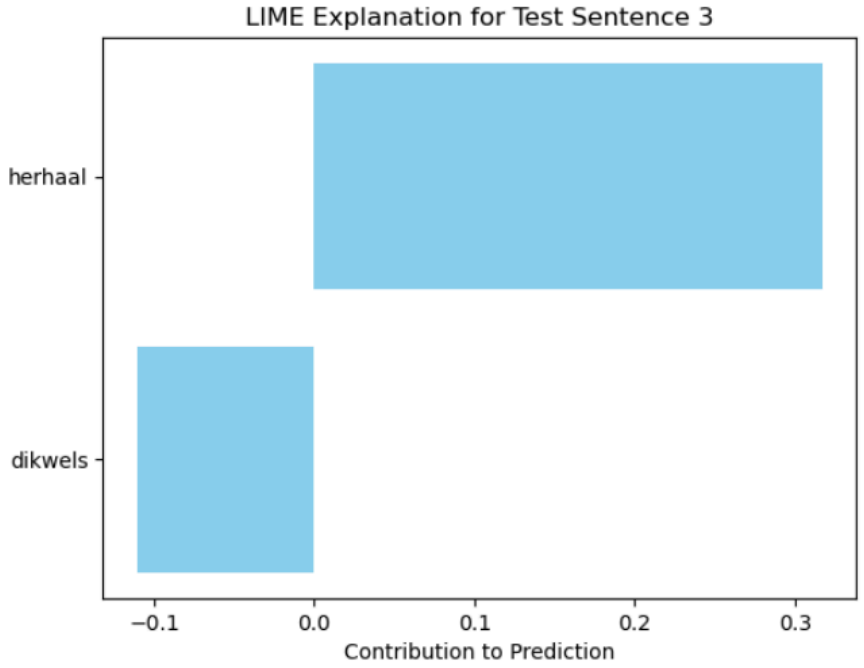

```
Test Sentence 1: vang 'n inbreker.
Actual Label: 0
Predicted Label: 1
Explanation: [('inbreker', 0.4651276018056682), ('vang', 0.10473695970484932), ('n', 0.03907466618606224)]
```

Figure 79. Afrikaans Sentence 1 Explanation



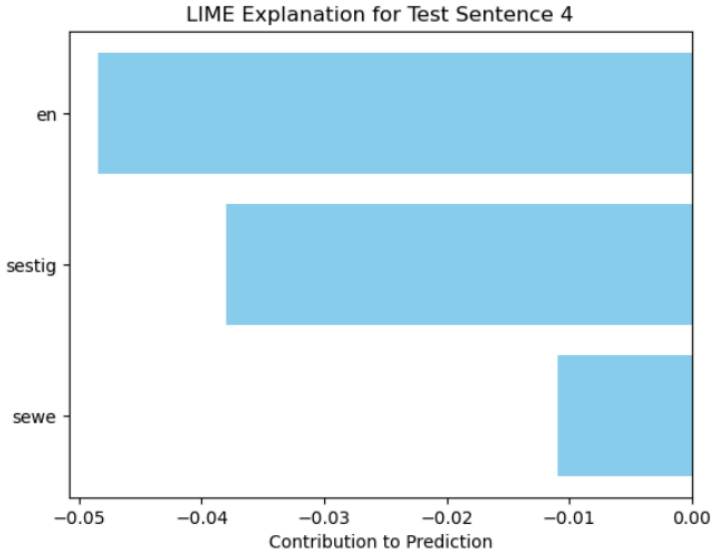

```
Test Sentence 2: leef sagkens.
Actual Label: 0
Predicted Label: 0
Explanation: [('leef', -0.019327724164419528), ('sagkens', 0.006805914481403187)]
```

Figure 80. Afrikaans Sentence 2 Explanation



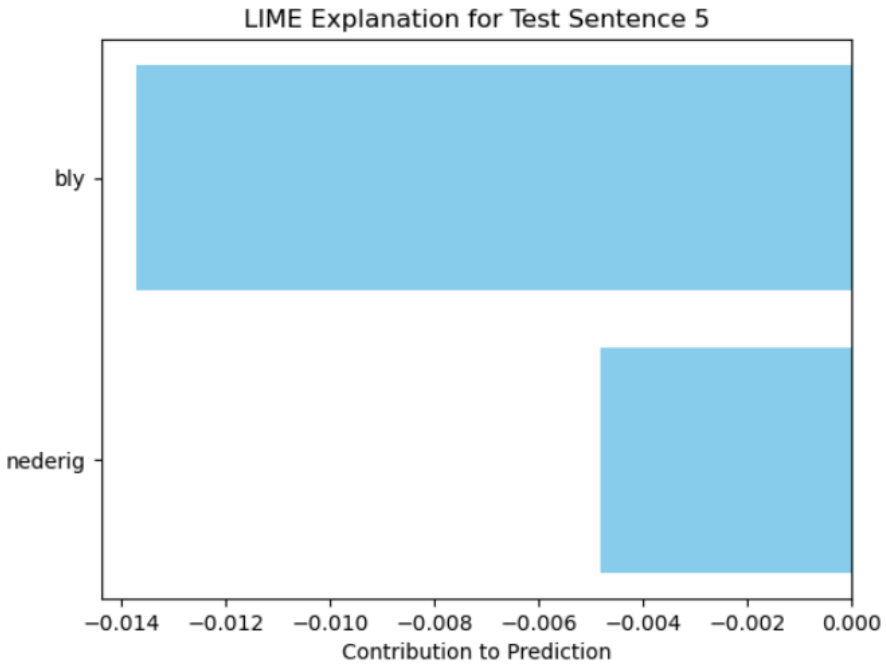

```
Test Sentence 3: herhaal dikwels.
Actual Label: 0
Predicted Label: 1
Explanation: [('herhaal', 0.3168544064583523), ('dikwels', -0.11045258771286758)]
```

Figure 81. Afrikaans Sentence 3 Explanation




```
Test Sentence 4: sewe-en-sestig probeerslae.
Actual Label: 0
Predicted Label: 2
Explanation: [('en', -0.04841465212064523), ('sestig', -0.038038525486237366), ('sewe', -0.010897593880906956)]
```

Figure 82. Afrikaans Sentence 4 Explanation




```
Test Sentence 5: bly nederig.
Actual Label: 0
Predicted Label: 0
Explanation: [('bly', -0.013708876264304259), ('nederig', -0.00480012613964212)]
```

Figure 83. Afrikaans Sentence 5 Explanation

5.3 Insights

5.3.1 Translation and sentiment analysis

The dataset “cleaned_dataset.xslx”, contained the translations and sentiment values across the multiple languages. The dataset contains translations of words including French, English, Afrikaans, Sepedi and IsiZulu. Along with each word’s associated sentiment score and provided sentiment. Once the data is loaded, the code performs some basic preparation steps. It renames columns to ensure they are consistently labelled, removes punctuation and converts all the language text to lowercase to ensure standardisation for the code to recognize words without being affected by variations in capitalization or punctuation.

After the dataset is cleaned, the code creates two dictionaries to facilitate translation and scoring. The first dictionary maps Choice A of words to the Choice B of translations. Meaning the program can take the choice of English words and translate them to the choice of IsiZulu words. This is determined by the user changing the inputs in the code for translation_lexique for Choice A and lexique for Choice B. The second dictionary is a scoring dictionary that associates each Choice B word with a sentiment score that indicates how positive, negative or neutral that word is.

To analyse the word sentiment, the code defines a function called analyse_sentiment. This function takes a sentence, splits it into individual words and looks up each word in the scoring dictionary. It then adds up the scores for all the words in the sentence to determine an overall sentiment score. If this score is above 0.05, the sentence is classified as positive (Positif). If the score is below -0.05, it is classified as negative (Négatif). Scores that fall between these values are marked as neutral (Neutre).

5.3.2 Creation and testing of sentence corpus

The testing corpus was created manually using English words extracted from the lexicon. Sentences were carefully constructed to carry either a positive or negative sentiment. This approach ensured that each sentence could effectively test the sentiment analysis functionality. After constructing the English sentences, translations were completed for the target languages. Finally, the sentiment scores were allocated to each translated sentence row. “Sentences dataset.xlsx" contains all the sentences translated from English, Afrikaans, Sepedi and IsiZulu.

The translation works by utilising the translate_text_using_lexicon function to translate an entire sentence by breaking it down into individual words. Each word is cleaned by removing any punctuation and converting it to lowercase. It then looks up each word in the lexicon for a translation. If a translation is found, it’s used. If not, the original word is kept. Once all words are translated, the function capitalises the first word of the sentence and ensures all other words are lowercase for readability. The result is a translated sentence that maintains a natural sentence structure. For each translated sentence the function analyse_sentiment evaluates the emotional tone of the sentence based on the translated words, calculating a total score that reflects whether the sentence is positive, negative, or neutral. The translated sentence, the total score (indicating sentiment strength), the overall sentiment classification (positive, negative, or neutral) and scores for individual words in the sentence are displayed as the results.

	Translated Text	Total Score	Sentiment
0	Inyakanyaka isihawu	-1	Négatif
1	Ubuwula inzondo	-10	Négatif
2	Shukuza nkosazana	8	Positif
3	Inzondo nkosazana	-3	Négatif
4	Isihawu shaya	4	Positif

	Word Scores
0	{'inyakanyaka': -3, 'isihawu': 2}
1	{'ubuwula': -3, 'inzondo': -7}
2	{'shukuza': 4, 'nkosazana': 4}
3	{'inzondo': -7, 'nkosazana': 4}
4	{'isihawu': 2, 'shaya': 2}

Figure 85. Results of translated text, total score, overall sentiment and word scores

5.3.3 Translation insights for the lexicon translating words from English to Sepedi

(Translation and Sentiment analysis functions)

In the expanded lexicon to translate English words into Sepedi, the functions fetch the translations from English to Sepedi which results in sentences in Sepedi that contain direct translations. The direct translations of these sentences do not account for grammar, context, and semantic meaning. An example would be in translation of proverbs from English to the Sepedi language. Since the lexicon does not account for the cultural context of these languages the translation of proverbs would generally not make sense.

Using the provided expanded lexicon the following English proverb was translated to Sepedi:

English sentence: “All good things must come to an end”.

Sepedi direct translation using given lexicon: “ka moka gabotse dilo must tla to an mafelelo”.

Correct translation: “Dilo ka moka tše dibotse di swanetše go fela”.

Firstly, it is important to note that the lexicon is limited in vocabulary, therefore the word “must” is not translated. Secondly, the lexicon did direct translations word-for-word resulting in a sentence that does not make complete grammatical sense in the Sepedi language, the correct grammar for Sepedi was ignored. Lastly, the Sepedi language has pronunciations and features that are not existent in the English language meaning the lexicon will fail to translate these “special words”.

The lexicon also has a limitation in translating homonyms, which are words that have the same spelling but different meanings. An example would be the word “light” – referring to a “light” as illumination or “light” as weight.

To improve the lexicon for translation from English to Sepedi language, the vocabulary needs to be extended, the lexicon must also be extended to a supervised learning algorithm that understands the grammar of Sepedi – allowing for correct grammatical translations. Lastly, the lexicon needs expansion to include extra Sepedi words that do not exist in English but have meaning in the context of a sentence formulated in English.

6 - CONCLUSION

After trying the two approaches being i) training the machine learning models to predict the nature and sentiment of word based on the expanded lexicon and ii) training the machine learning models to predict the sentiment of a sentence based on the new created corpus, one can conclude that the model's performances do not deviate that much from each other. All models have an accuracy that is lower than 70%. Which means that they performed in the same range, this can be due to a few reason being i) the low accuracy of ML trained on expanded lexicon were faced with data that was either not 100 % cleaned or the dataset was too complex for the models ii) the ML trained on the corpus may have also performed low because the sentences used where derived from the expanded lexicon thus the two dataset are that different from each other. Recommendations when dealing with lexicons may be to allow more time into preprocessing your dataset and understanding its shortcomings.

7 - REFERENCES

Aasir, M. (2024) 'Unveiling Insights from Amazon Fine Food Reviews: A Data-Driven Exploration', *Medium*. Available at: https://medium.com/@mohamedaasir1992/unveiling-insights-from-amazon-fine-food-reviews-a-data-driven-exploration-caf13092cd88

Abdullah, N.A.S. & Rusli, N.I.A. 2021. Multilingual Sentiment Analysis: A Systematic Literature Review. Pertanika Journal of Science & Technology, 29(1).

Agüero-Torales, M.M., Abreu Salas, J.I. & López-Herrera, A.G. 2021. Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, 107:107373.

Araújo, M., Pereira, A. & Benevenuto, F. 2020. A comparative study of machine translation for multilingual sentence-level sentiment analysis. Information Sciences, 512:1078-1102.

Barbosa, M.W. & Gomes, A. 2025. Themes and sentiments in conversations about food waste on Twitter: Proposal of a framework using neural topic modelling. Food Quality and Preference, 122:105311.

Dervenis, C., Kanakis, G. & Fitsilis, P. 2024. Sentiment analysis of student feedback: A comparative study employing lexicon and machine learning techniques. Studies in Educational Evaluation, 83:101406.

Farhoudinia, B., Ozturkcan, S. & Kasap, N. 2024. Emotions unveiled: detecting COVID-19 fake news on social media. Humanities and Social Sciences Communications, 11(1):640.

Kumari, S. & Singh, M.P. 2024. Machine Learning-Based Election Results Prediction Using Twitter Activity. SN Computer Science, 5(7).

Mabokela, K.R., & Schlippe, T. (2022). 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.

So, C. 2021. Understanding the Prediction Mechanism of Sentiments by XAI Visualization. Paper presented at Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, Seoul, Republic of Korea:75–80. [Online]. Available from: https://doi.org/10.1145/3443279.3443284.

Suresh, H. (2020) 'Wordclouds & Basics of NLP', *Medium*. Available at: https://medium.com/@harinisureshla/wordclouds-basics-of-nlp-5b60be226414