Sentiment Analysis for the Mizo Language: A Comparative Study of Classical Machine Learning and Transfer Learning Approaches

Mercy Lalthangmawii and **Thoudam Doren Singh**
National Insitute of Techology Silchar, India
mercy_pg_22@cse.nits.ac.in, doren@cse.nits.ac.in

Abstract

Sentiment analysis, a subfield of natural language processing (NLP) has witnessed significant advancements in the analysis of user-generated contents across diverse languages. However, its application to low-resource languages remains a challenge. This research addresses this gap by conducting a comprehensive sentiment analysis experiment in the context of the Mizo language, a low-resource language predominantly spoken in the Indian state of Mizoram and neighboring regions. Our study encompasses the evaluation of various machine learning models including Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbor (K-NN), Logistic Regression and transfer learning using XLM-RoBERTa. The findings reveal the suitability of SVM as a robust performer in Mizo sentiment analysis demonstrating the highest F-1 Score and Accuracy among the models tested. XLM-RoBERTa, a transfer learning model exhibits competitive performance highlighting the potential of leveraging pre-trained multilingual models in low-resource language sentiment analysis tasks. This research advances our understanding of sentiment analysis in low-resource languages and serves as a stepping stone for future investigations in this domain.

1 Introduction

The ever-growing digital landscape has given rise to an abundance of user-generated content across various online platforms. This proliferation of text data, often accompanied by valuable insights and opinions, has sparked significant interest in the field of sentiment analysis. Sentiment analysis, a subset of natural language processing (NLP), seeks to discern the emotional tone and opinions expressed within textual content (Pang et al., 2002). Its applications span from understanding customer sentiment for business intelligence (Liu, 2012) to monitoring public sentiment in the political sphere (O'Connor et al., 2010).

While sentiment analysis has seen considerable progress in many widely spoken languages, it has encountered substantial challenges when applied to low-resource languages such as the Mizo language. The Mizo language, predominantly spoken in the Indian state of Mizoram and neighboring regions belongs to the Kuki-Chin branch of the Sino-Tibetan language family. It holds cultural and regional significance yet remains relatively underrepresented in the domain of computational linguistics. The Mizo language serves as the gateway to understanding the sentiments and opinions of the Mizo-speaking community in their native tongue and unlocking this holds immense potential for diverse applications including social media monitoring, community feedback analysis and local sentiment-driven decision-making.

In addressing the compelling need to develop sentiment analysis models tailored to low-resource languages like Mizo, it is crucial to delve into some linguistic details. The Mizo language being part of the Kuki-Chin branch shares linguistic affiliations with languages like Hmar, Lushai (Mizo), and Paite. The Mizo language also diverges from English not only in structure and word order but in its declarative sentence construction which follows an Object-Subject-Verb (OSV) order. In its basic or unmarked word order, Mizo tends to build sentences in a way where the main elements such as the subject or object come first on the left side and additional information is added to the right. Instead of using prepositions (like "under" or "over"), Mizo uses postpositions which are similar but come after the noun they modify. For example, instead of saying "on the table", in Mizo, we say "table chungah" which literally translates to "the table on". The positioning of auxiliary verbs after the main verb is a distinctive feature as it contrasts with the typical positioning found in many other languages; consider the English sentence "She has completed the homework" which translates to "Homework a zo tawh". Here, "tawh" is an auxiliary verb indicating completion or emphasis. Adverbs are also mainly placed at the end of the sentence after the verb, eg : "I am not going" will be "Ka kal dawn lo" where "lo" signifies the negation. Mizo has a sophisticated system of agreement in which various parts of a sentence (like subjects and objects) can have matching features. This complexity allows for the omission or dropping of pronouns in certain cases, known as 'heavy pro drop.' For eg: "He sings a song and everyone loves him" can be written as "Hla a sa a, an duh hle"; here the pronoun for him is not included in the Mizo version due to Mizo's rich agreement patterns and the ability to mark features on verbs that align with subjects and objects.

2 Literature Survey

Sentiment analysis in natural language processing (NLP) has emerged as a dynamic field, driven by the need to comprehend the emotional tone and attitudes expressed in the ever-expanding realm of user-generated content across diverse platforms (Nguyen et al., 2021). The applications of sentiment analysis span various domains including business brand surveys, social media studies, and movie analyses (Pérez et al., 2021). This process involves categorizing opinions into positive, negative, or unbiased classes offering valuable insights into product popularity and market dynamics (Parikh and Shah, 2021).

In response to the growing demand for sentiment analysis tools, researchers have developed resources like pysentimiento, a library facilitating sentiment analysis in multiple languages, such as Spanish and English (Yue et al., 2019). The advancement of deep learning models including long-short term memory and attention models has significantly improved the accuracy of sentiment analysis tasks (Devika et al., 2016; Panthati et al., 2018; Shilpa et al., 2021). Additionally, lexicon-based approaches and traditional machine learning techniques like Naive Bayes and Support Vector Machines (SVM) continue to be widely employed in sentiment analysis (Umarani et al., 2021; Ahmad Hapez et al., 2021; Khamphakdee et al., 2021; Tammina, 2020).

While sentiment analysis has made substantial progress in well-represented languages, research in low-resource languages has been comparatively limited due to language complexity and resource constraints. However, recent studies have addressed this gap, contributing to the exploration of sentiment analysis in linguistically diverse and underrepresented languages. For instance, in the case of Manipuri language, the authors (Meetei et al., 2021) perform language-specific pre-processing tasks such as transliteration building negative morpheme-based lexicon and filtering noisy words which improve classification results. Some of the other low resource language works on sentiment classification of the social media contents are also found in (Das and Singh, 2021, 2023; Singh et al., 2021). In the context of code-mixed languages, (Roy, 2023) proposed an ensemble model for predicting sentiment in code-mixed Kannada and Malayalam achieving promising results. In the realm of low-resource African languages, (Wang et al., 2023a) highlighted the significance of adapting pretrained models to target languages and tasks demonstrating significant performance improvements. (Hameed et al., 2023a) explored classical machine learning and neural network-based techniques for sentiment analysis of Central Kurdish emphasizing the effectiveness of data augmentation.

Moreover, research on sentiment analysis in Roman Urdu by (Ullah et al., 2022) involved the creation of a dedicated corpus and the application of various models with the CNN-LSTM method yielding the most favorable results. The study by (Wang et al., 2022) proposed DLAN, an end-to-end sentiment analysis architecture utilizing adversarial learning and knowledge distillation for cross-lingual sentiment analysis showcasing improved performance compared to baseline methods. (Das and Singh, 2021) proposed a model that uses lexical features like adjectives, adverbs, and verbs to classify the sentiment polarity of Assamese news article and their model showed improvement over the baseline model in terms of F1-score on a standard dataset.

While acknowledging these contributions, this paper aims to extend the discourse by addressing sentiment analysis in the Mizo language.

3 Methodology

In this section, we outline the methodology employed for collecting, preprocessing, and annotating the datasets used in our sentiment analysis research with a focus on data collection, compilation, and the labeling process.

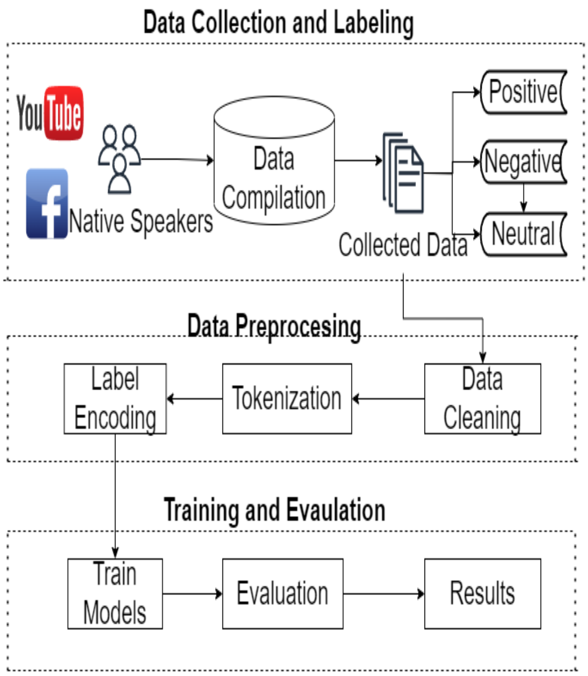

Figure 1: Workflow for Mizo Sentiment Analysis

3.1 Data collection

Our dataset was sourced from social media platforms primarily YouTube and Facebook where user-generated comments provides a rich source of textual data for sentiment analysis. Given the low-resource nature of the Mizo language, gathering data directly from native speakers was crucial to ensure the authenticity and relevance of the dataset. Native speakers were encouraged to contribute their comments and opinions fostering a dataset that is representative of authentic sentiments expressed in the Mizo language. These contributions primarily consist of textual data from short day-to-day conversations reflecting the informal and colloquial expressions commonly used in text communication. The dataset primarily comprises of tweet-level granularity with an average character count falling within the range of 35 to 100 characters.

The number of dataset along with their overall word count for each of the sentiments are given in table 1.

Table 1: Number of dataset with word count for each sentiment

S.No	Categories	Dataset Count	Word Count
1.	Positive	2768	16714
2.	Negative	2042	22519
3.	Neutral	1400	10661

3.2 Data Labeling

The core of our dataset preparation involved labeling each comment with its corresponding sentiment category: positive, negative or neutral. Each comment in the dataset was manually examined by annotators who possessed a strong understanding of the Mizo language and cultural context. Annotators assigned sentiment labels based on the overall emotional tone and attitude expressed in the comments according to their opinions. A 'neutral' category was included to capture comments that did not exhibit a discernible positive or negative sentiment. These comments were carefully categorized to account for a wide range of expressions. To enhance the robustness and reliability of the labeling process, comments were reviewed and labeled by three separate annotators independently. In cases where discrepancies arose, the sentiment label assigned to each comment was determined by majority consensus. To assess the agreement among annotators we employed Fleiss's Kappa as a metric to quantify the level of agreement among our three annotators from table 2. Fleiss's Kappa was chosen as the agreement metric due to its applicability to situations with more than two raters. The formula for Fleiss's Kappa is as follows:

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

Where: κ is Fleiss's Kappa, P_o is observed agreement, and P_e is expected agreement.

The calculated Fleiss's Kappa for our dataset yielded a value of approximately 0.88. This substantial agreement value indicates a high level of concordance among annotators, suggesting a consistent understanding and application of annotation criteria across the three raters.

Table 2: Annotation Agreement

	Annotator A	Annotator B	Annotator C
Negative	2015	2020	2035
Positive	2775	2765	2770
Neutral	1420	1415	1400

3.3 Preprocessing

In this section, we detail the preprocessing steps undertaken to cleanse and optimize the raw textual data extracted from the social media platforms and native speaker contributions. Preprocessing is a critical phase in data preparation for sentiment analysis as it ensures that the text data is in a suitable format for subsequent analysis and modeling.

To standardize and clean the text data, we implemented a series of steps:

- Lowercasing : All text was converted to lowercase to ensure uniformity in text representation.
- Whitespace Removal: Extraneous whitespaces, newline characters and carriage returns were removed to enhance data consistency.
- Special Character Removal: We employed regular expressions to eliminate special characters and symbols from the text retaining only alphanumeric characters and spaces.

Data integrity is of paramount importance. To ensure the highest quality dataset, any instances with missing or null values were removed from consideration. This step further enhanced the reliability and consistency of the dataset. Samples of the dataset are shown in Fig 2.

3.3.1 Label Encoding

In order to facilitate sentiment analysis, it was necessary to encode the sentiment labels into numerical values. We employed label encoding techniques to map sentiment categories, such as 'negative','positive,' 'neutral' to corresponding numeric labels. Table 3 shows the label encoding for the sentiments. This encoding process enabled us to feed the labeled data into machine learning models for sentiment classification.

Table 3: Label encoding for the sentiments

S.No	Categories	Label Encoding
1.	Negative	0
2.	Neutral	1
3.	Positive	2

3.4 Classification Algorithms

We experiment with a total of five machine learning models which are commonly used for sentiment analysis. In this work, the logistic regression, support vector machine, decision tree, random forest and K-nearest neighbour algorithms are considered for the experiments.

Index	emotion	text
0	negative	A va rapthlak ve
1	negative	Ava mak ve he nu hi kan Veng nu ani hi
2	positive	a hlawk dawn rem rem
3	neutral	A va hlaawm ve
4	neutral	A part thenkhat te hi an lo ruk sak lovang tih pawh a sawi theih loh chu a
5	negative	A vahan rapthlak ngai em atider tute hi ale…
6	negative	Avuakhlum tur nih khi a ma lungpawp lo kha tih pawp sak zawk tlak a nih

Figure 2: Samples from the collected data

Algorithm 1 Sentiment Analysis using Classification Algorithms

Input: Raw textual data (T)
Output: Positive/Negative/Neutral

procedure DATA PREPROCESSING
    $Token_i \leftarrow \text{clean\_text}(T)$
    $I_c \leftarrow \text{Preprocess\_text}(Token_i)$ ▷ Token-level text preprocessing
    $L_c \leftarrow \text{Label\_encoding}(Token_i)$ ▷ Encoding labels for sentiments
    $Fv_i \leftarrow \text{Feature\_Engineering}(Fv_i)$ ▷ Using TF-IDF for feature extraction
procedure MODEL TRAINING
    $X\_train, X\_test \leftarrow \text{train\_test\_split}(\text{I}_c)$ ▷ 80/20 split
    $Model\_Train(\text{input} = Fv_i, \text{Polarity})$ ▷ Training the sentiment analysis model
    Trained_Model
procedure MODEL PREDICTION
    Polarity $\leftarrow$ Positive/Negative/Neutral
    $X\_test \leftarrow \text{preprocessed\_text}(T_c)$
    $Pp \leftarrow \text{Trained\_Model.predict}(X\_Test)$ ▷ Using the trained model for prediction
    Return *Pp*

3.5 Transfer Learning

Transfer learning has emerged as a transformative technique in the field of sentiment analysis particularly in the context of low-resource languages. This approach leverages pre-existing knowledge gained from high-resource languages and datasets to enhance sentiment analysis performance in languages with limited linguistic resources. By transferring the knowledge encoded in pretrained language models, transfer learning offers a valuable shortcut to adapt sentiment analysis models for new linguistic environments.

Transfer learning encompasses various strategies including fine-tuning and feature extraction, to repurpose pretrained models for sentiment analysis tasks. Researchers have explored the adaptability of these strategies in cross-lingual and cross-domain sentiment analysis scenarios achieving promising results.(Hameed et al., 2023b) describe the collection and annotation of a dataset for sentiment analysis of Central Kurdish and demonstrate the effectiveness of data augmentation using pretrained models.(Wang et al., 2023b) propose language-adaptive and task-adaptive pretraining on African texts for sentiment analysis in low-resource African languages achieving remarkable performance improvements.(Mamta et al., 2022) propose a deep multi-task multi-lingual adversarial framework that leverages knowledge from a high-resource language to solve the resource-scarcity problem in sentiment analysis achieving significant performance gains.(de Toledo and Marcacini, 2022) introduce a transfer learning approach using joint fine-tuning for sentiment analysis in multimodal data, providing competitive results with efficient computation.(Nugumanova et al., 2022) demonstrate the effectiveness of transfer learning using pretrained multilingual models for sentiment analysis in low-resource languages with improved performance after fine-tuning .

Algorithm 2 Sentiment Analysis using XLM-RoBERTa

Input: Raw textual data (T)
Output: Positive/Negative/Neutral

procedure DATA PREPROCESSING
  $\text{Token}_i \leftarrow \text{clean\_text}(T)$
  $I_c \leftarrow \text{Preprocess\_text}(Token_i)$
procedure MODEL TRAINING
  $X\_tr, X\_te \leftarrow \text{train\_test\_split}(\text{I}_c)$ ▷ 80/20 split
  $\text{T}_k \leftarrow \text{XLMRobertaTokenizer}(\text{X\_tr}, \text{X\_te})$
  model ← xlm-roberta-base
  num_epochs ← 50
  $Model\_Train(\text{T}_k)$
  Trained_Model
procedure MODEL PREDICTION
  Polarity ← Positive/Negative/Neutral
  $X\_test \leftarrow \text{preprocessed\_text}(\text{T}_k)$
  $Pp \leftarrow \text{Trained\_Model.predict}(X\_test)$ ▷ Using the trained model for prediction
  Return *Pp*

In our proposed model, we explore transfer learning using XLM-RoBERTa, a state-of-the-art pretrained model without fine-tuning. This approach allows us to capitalize on the general language understanding encoded in the model while tailoring it specifically to the Mizo language. By incorporating transfer learning into our sentiment analysis framework, we aim to shed light on its applicability and effectiveness in addressing the challenges posed by low-resource languages contributing to the growing body of knowledge in this domain.

Algorithm 3 Sentiment Analysis using BERT

Input: Raw textual data (T)
Output: Positive/Negative/Neutral

procedure DATA PREPROCESSING
  $\text{Token}_i \leftarrow \text{clean\_text}(T)$
  $\text{I}_c \leftarrow \text{Preprocess\_text}(\text{Token}_i)$
procedure MODEL TRAINING
  $Classifier \leftarrow$ BERT
  $\text{C}_l \leftarrow \text{Candidate\_labels}$ ▷ Postive/Negative/Neutral
  $Classification \leftarrow Classifier(\text{I}_c, \text{C}_l)$
procedure MODEL PREDICTION
  Input: Preprocessed data and candidate labels
  Output: Predicted sentiment label
  $Pp \leftarrow \text{Predicted\_labels}(Classification)$
  Return *Pp*

3.6 Zero Shot Learning

Zero-shot classification is a paradigm in machine learning and natural language processing that enables the classification of data points into categories that the model has never seen during training. By utilizing zero-shot techniques, a pre-trained multilingual model can categorize text in the low-resource language into their corresponding sentiments based on its understanding of sentiment patterns and language semantics gleaned from its exposure to other languages during pre-training. Several notable studies have delved into the effectiveness of leveraging pre-trained multilingual models such as mBERT and XLM-R, for sentiment analysis in low-resource languages (Van Thin et al., 2023; Yang et al., 2022). These investigations have yielded promising outcomes, highlighting that fine-tuning multilingual models can achieve commendable performance in zero-shot cross-lingual scenarios (Phan et al., 2021). Moreover, research indicates that selecting source languages other than

English can be particularly advantageous when conducting sentiment analysis in low-resource language contexts (Kumar and Albuquerque, 2021). In this context, we explore zero-shot cross-lingual transfer learning with the BERT model and aim to demonstrate the applicability of this approach in the context of sentiment analysis for the Mizo language emphasizing its potential as a viable solution for low-resource languages.

4 Results and Discussions

Table 4: Performance of different models tested

Model	Categories	Precision	Recall	F-1 Score	Accuracy
Random Forest	Positive Negative Neutral	0.73 0.73 0.62	0.66 0.81 0.55	0.69 0.74 0.58	0.70
Support Vector Machine	Positive Negative Neutral	0.80 0.77 0.65	0.75 0.85 0.56	0.77 0.80 0.60	0.75
Decision Tree	Positive Negative Neutral	0.64 0.67 0.48	0.62 0.69 0.47	0.63 0.68 0.47	0.61
K-Nearest Neighbour	Positive Negative Neutral	0.75 0.73 0.55	0.67 0.82 0.50	0.71 0.77 0.53	0.69
Logistic Regression	Positive Negative Neutral	0.77 0.78 0.61	0.76 0.83 0.54	0.76 0.81 0.57	0.74
XLMR-Roberta	Positive Negative Neutral	0.76 0.78 0.68	0.80 0.84 0.54	0.78 0.81 0.60	0.75
BERT	Positive Negative Neutral	0.38 0.46 0.25	0.13 0.89 0.03	0.19 0.61 0.05	0.45

As per the findings of our experiment from Table 4, several key insights can be observed. One of the standout findings of our experiment is the high performance of the Support Vector Machine (SVM) model across all sentiment categories. It achieved the highest F-1 Score and Accuracy among all models tested. XLMR-Roberta demonstrated competitive performance particularly in terms of Precision and Recall maintaining high Precision values while achieving a respectable Recall rate. This indicates that transfer learning using pre-trained multilingual models can be a viable approach for low-resource language sentiment analysis.The other classic ma-

Table 5: Error Rates of each Model

S.No	Model Name	Sentiments	Error Rate
1.	Support Vector Machine	Neutral Positive Negative	0.439 0.247 0.152
2.	Random Forest	Neutral Positive Negative	0.494 0.318 0.177
3.	Decision Tree	Neutral Positive Negative	0.527 0.362 0.314
4.	K-Nearest Neighbour	Neutral Positive Negative	0.498 0.333 0.182
5.	Logistic Regression	Neutral Positive Negative	0.442 0.254 0.172
6.	XLM-RoBERTa	Neutral Positive Negative	0.525 0.186 0.134
7.	BERT	Neutral Positive Negative	0.972 0.878 0.105

chine learning models delivered moderate performances with reasonable metrics. In our experimentation, the BERT model demonstrated suboptimal performance; it exhibited diminished Precision and Recall metrics particularly concerning Positive and Neutral sentiment categories culminating in a reduced F-1 Score and Accuracy. This observed performance inadequacy can be attributed to the unique evaluation setting wherein the model underwent assessment in a zero-shot classification paradigm. This signifies that the model was not provided with dedicated training and testing sets tailored to the nuances of sentiment analysis in the Mizo language.

The disparities evident in the model predictions can be ascribed to several pivotal factors, primarily rooted in the constraints of the training dataset. Notably, the significant misclassification rates within the 'neutral' sentiment category as delineated in Table 5 underscore its inadequate representation within the dataset. This imbalance poses a substantial challenge to the model impeding its capacity to discern and classify instances accurately within the 'neutral' sentiment category. An intriguing avenue for further exploration lies in augmenting the sentiments in each category potentially alleviating the impact of dataset imbalances on model performance. A prospective analysis of how these metrics evolve with an increased representation of senti-

Input Text for Mizo	Translation to English
mahni in mawhpulh ringawt tur ani ngai lo e mahni inhmanghaih hi a tha a nia	it's not good to blame yourself instead you should love yourself.
kan in biak a va trul ve	we really need to talk
a va rapthlak em hmeichhia leh nghal	that is horrible for a a woman
bengvarna chauh hi mi an hrilh thin aa midang te hrilh zawk ulaaaa rip	you only tell me what is popular, go tell others
tang fan fan mizo hnahthlak te u	full support for the mizos
min hrelo der chu ka lawmlo ania aw	I am not happy you pretend to not know me
bang poh ni se in pha lo buaina mai ani	even if they resign, it is useless
nula chuan pangpar dawrah hna thawh nuam a ti hle	she really likes to work at the flower shop
ka pan min hmangaih thu a ziak	my father is expressing his genuine love

Figure 3: Mizo - English Translation of the Text Inputs for Machine Learning Models

Input_Text for Random Forest	Predicted_Sentiment_RF	Actual_Sentiment
mahni in mawhpulh ringawt tur ani ngai lo e mahni inhmanghaih hi a tha a nia	positive	positive
kan in biak a va trul ve	negative	neutral
a va rapthlak em hmeichhia leh nghal	negative	negative
bengvarna chauh hi mi an hrilh thin aa midang te hrilh zawk ulaaaa rip	negative	negative
tang fan fan mizo hnahthlak te u	neutral	positive
min hrelo der chu ka lawmlo ania aw	negative	negative
bang poh ni se in pha lo buaina mai ani	neutral	neutral
nula chuan pangpar dawrah hna thawh nuam a ti hle	positive	positive
ka pan min hmangaih thu a ziak	negative	positive
Input_Text for Support Vector Machine	**Predicted_Sentiment_SVM**	**Actual_Sentiment**
mahni in mawhpulh ringawt tur ani ngai lo e mahni inhmanghaih hi a tha a nia	negative	positive
kan in biak a va trul ve	negative	neutral
a va rapthlak em hmeichhia leh nghal	negative	negative
bengvarna chauh hi mi an hrilh thin aa midang te hrilh zawk ulaaaa rip	negative	negative
tang fan fan mizo hnahthlak te u	neutral	positive
min hrelo der chu ka lawmlo ania aw	negative	negative
bang poh ni se in pha lo buaina mai ani	neutral	neutral
nula chuan pangpar dawrah hna thawh nuam a ti hle	positive	positive
ka pan min hmangaih thu a ziak	positive	positive
Input_Text for Decision Tree	**Predicted_Sentiment_DT**	**Actual_Sentiment**
mahni in mawhpulh ringawt tur ani ngai lo e mahni inhmanghaih hi a tha a nia	positive	positive
kan in biak a va trul ve	negative	neutral
a va rapthlak em hmeichhia leh nghal	negative	negative
bengvarna chauh hi mi an hrilh thin aa midang te hrilh zawk ulaaaa rip	negative	negative
tang fan fan mizo hnahthlak te u	neutral	positive
min hrelo der chu ka lawmlo ania aw	negative	negative
bang poh ni se in pha lo buaina mai ani	neutral	neutral
nula chuan pangpar dawrah hna thawh nuam a ti hle	positive	positive
ka pan min hmangaih thu a ziak	positive	positive
Input_Text for K-Nearest Neighbour	**Predicted_Sentiment_KNN**	**Actual_Sentiment**
mahni in mawhpulh ringawt tur ani ngai lo e mahni inhmanghaih hi a tha a nia	negative	positive
kan in biak a va trul ve	negative	neutral
a va rapthlak em hmeichhia leh nghal	negative	negative
bengvarna chauh hi mi an hrilh thin aa midang te hrilh zawk ulaaaa rip	negative	negative
tang fan fan mizo hnahthlak te u	negative	positive
min hrelo der chu ka lawmlo ania aw	negative	negative
bang poh ni se in pha lo buaina mai ani	neutral	neutral
nula chuan pangpar dawrah hna thawh nuam a ti hle	positive	positive
ka pan min hmangaih thu a ziak	positive	positive
Input_Text for Logistic Regression	**Predicted_Sentiment_LogReg**	**Actual_Sentiment**
mahni in mawhpulh ringawt tur ani ngai lo e mahni inhmanghaih hi a tha a nia	negative	positive
kan in biak a va trul ve	negative	neutral
a va rapthlak em hmeichhia leh nghal	negative	negative
bengvarna chauh hi mi an hrilh thin aa midang te hrilh zawk ulaaaa rip	negative	negative
tang fan fan mizo hnahthlak te u	neutral	positive
min hrelo der chu ka lawmlo ania aw	negative	negative
bang poh ni se in pha lo buaina mai ani	neutral	neutral
nula chuan pangpar dawrah hna thawh nuam a ti hle	positive	positive
ka pan min hmangaih thu a ziak	positive	positive

Figure 4: Text Inputs and Predictions for Machine Learning Models

Input Text for Mizo	Translation to English	Actual_Label	Predicted_Label	Predicted_Score
a va rapthlak ve	that is horrible	negative	negative	0.53
ava mak ve he nu hi kan veng nu ani hi	this is totally shocking, she lives in my area	negative	negative	0.877
a hlawk dawn rem rem	that's gonna be very fortunate	positive	negative	0.753
a va hlaawm ve	that is too far	neutral	negative	0.544
a part thenkhat te hi an lo ruk sak lovang tih pawh a sawi theih loh chu a	how can you tell they won't steal another part	neutral	negative	0.408
a vahan rapthlak ngai em atider tute hi aletling zawng a khai mai chi an ni lo maw kakhawngai hlawm emai	that is terrible, I'm feeling so sorry	negative	negative	0.586
hriat chian loh hi engmah deal neih pui loh tur	don't do any deal that you're not well-informed	neutral	negative	0.512
police ah thlenmai ula	you should go to the police	neutral	negative	0.772
nuaithum sem angvel anih hi	it's like distributing 3 lakhs	neutral	neutral	0.585

Figure 5: Text Inputs and Predictions for BERT

Input_Text for Mizo	Translation to English	Predicted_Sentiment	Actual_Sentiment
a nun zawng zawng tichhe tu i nih hi	that is the destroyer of life	Negative	Negative
hal vek ru	burn them all	Negative	Negative
ka unaunuin thil tui deuh min lei alawm	my sister bought me a tasty treat	Positive	Positive
ka hmu e	i can see it	Neutral	Neutral
an fel hlawm khawp mai	they are very friendly	Positive	Positive
hetah hian i park thei ang	you can park here	Neutral	Neutral
chawhnu lamah chuan hmuh che pawh ka lo inbeisei hleinem	I didn't expect to see you this afternoon	Neutral	Neutral
khawngaih takin hemi video hi lo en ve the u	you can watch the video here	Negative	Neutral
hmeichhe hmeltha tak ka tawng alawm le	i met a beautiful woman	Negative	Positive

Figure 6: Text Inputs and Predictions for XLM-RoBERTa

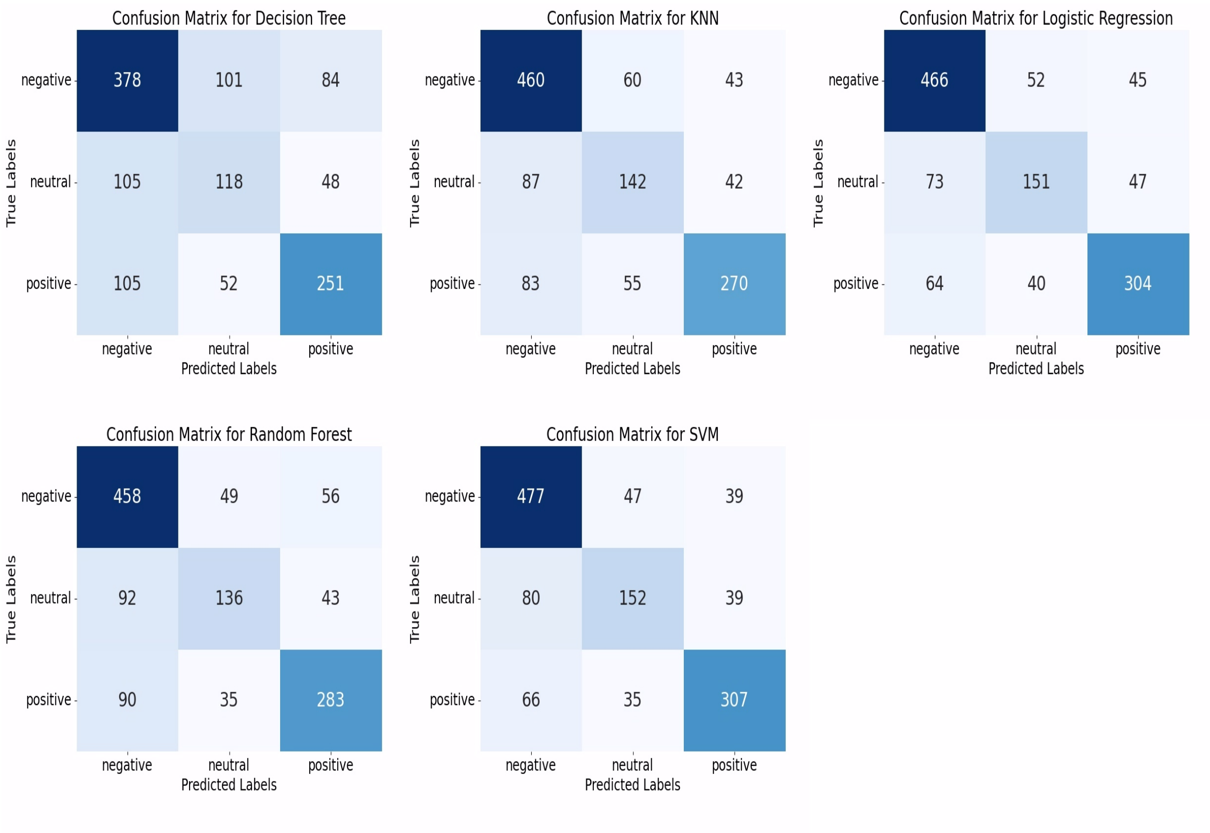

Figure 7: Confusion matrices of classical Machine Learning models

ments within each category holds the promise of providing valuable insights into model behavior.To visually capture the outcomes of the experiments, the resulting text inputs and model predictions are depicted from Figures 4 - 6, complemented by their respective confusion matrices from Figure 7-9.

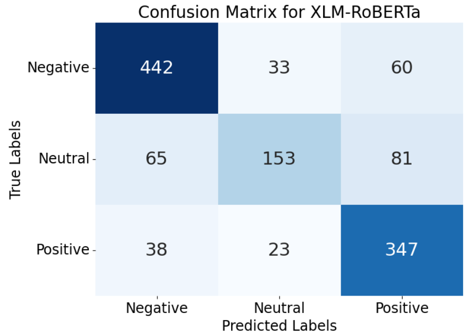

Figure 8: Confusion Matrix for XLM-RoBERTa

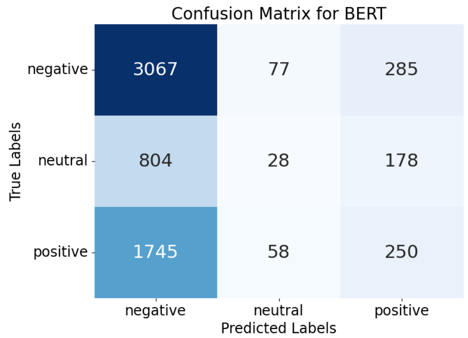

Figure 9: Confusion Matrix for BERT

5 Conclusion and Future Directions

This study contributes to the understanding of sentiment analysis in a low-resource language called "Mizo" and sheds light on the suitability of various machine learning models and approaches. SVM emerges as a strong performer, while transfer learning with XLMR-Roberta offers a promising avenue for enhancing sentiment analysis capabilities in resource-scarce language contexts.

Future studies should prioritize the development of more comprehensive and balanced datasets employing effective data augmentation and sampling techniques to rectify imbalances and improve the models' ability to generalize across various sentiment categories. Implementing advanced model training techniques including hyperparameter tuning and regularization strategies will be also pivotal in optimizing the models' learning processes and improving their predictive accuracy across a diverse range of sentiment inputs. By heeding these recommendations, future research endeavors can build upon the insights gained from this study and contribute to the development of more effective and reliable sentiment analysis methodologies for the Mizo language.

References

Muhamad Hariz Izzat Ahmad Hapez, Noor Latiffah Adam, and Zaidah Ibrahim. 2021. Performance analysis of machine learning techniques for sentiment analysis. In *Advances in Visual Informatics: 7th International Visual Informatics Conference, IVIC 2021, Kajang, Malaysia, November 23–25, 2021, Proceedings 7*, pages 205–213. Springer.

Ringki Das and Thoudam Doren Singh. 2021. A step towards sentiment analysis of assamese news articles using lexical features. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, pages 15–23. Springer.

Ringki Das and Thoudam Doren Singh. 2023. A hybrid fusion-based machine learning framework to improve sentiment prediction of assamese in low resource setting. *Multimedia Tools and Applications*.

Guilherme Lourenço de Toledo and Ricardo Marcondes Marcacini. 2022. Transfer learning with joint fine-tuning for multimodal sentiment analysis. *arXiv preprint arXiv:2210.05790*.

M D Devika, C Sunitha, and Amal Ganesh. 2016. Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, 87:44–49.

Razhan Hameed, Sina Ahmadi, and Fatemeh Daneshfar. 2023a. Transfer learning for low-resource sentiment analysis. *arXiv preprint arXiv:2304.04703*.

Razhan Hameed, Sina Ahmadi, and Fatemeh Daneshfar. 2023b. Transfer learning for low-resource sentiment analysis. *arXiv preprint arXiv:2304.04703*.

Nattawat Khamphakdee, Pusadee Seresangtakul, et al. 2021. Sentiment analysis for thai language in hotel domain using machine learning algorithms. *Acta Informatica Pragensia*, 10(2):155–171.

Akshi Kumar and Victor Hugo C Albuquerque. 2021. Sentiment analysis using xlm-r transformer and zero-shot transfer learning on resource-poor indian language. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–13.

B Liu. 2012. Sentiment analysis and opinion mining. synthesis lectures on human language technologies, 5 (1), 1-167.

Mamta, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–19.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, Samir Kumar Borgohain, and Sivaji Bandyopadhyay. 2021. Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation*, 55(4):947–969.

Viet Yen Nguyen, Andréa B Duque, Jules Belveze, Lukas D Baker, Astrid Harsaae, Pedro Pessoa, and Inês Oliveira. 2021. Atlastic reputation ai: Four years of advancing and applying a sota nlp classifier. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–2. IEEE.

Aliya Nugumanova, Yerzhan Baiburin, and Yermek Alimzhanov. 2022. Sentiment analysis of reviews in kazakh with transfer learning techniques. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–6. IEEE.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 122–129.

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

Jagadeesh Panthati, Jasmine Bhaskar, Tarun Kumar Ranga, and Manish Reddy Challa. 2018. Sentiment analysis of product reviews using deep learning. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2408–2414. IEEE.

Satyen M Parikh and Mitali K Shah. 2021. Analysis of various sentiment analysis techniques of nlp. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 1323–1331. IEEE.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462*.

Khoa Thi-Kim Phan, Duong Ngoc Hao, Dang Van Thin, and Ngan Luu-Thuy Nguyen. 2021. Exploring zero-shot cross-lingual aspect-based sentiment analysis using pre-trained multilingual language models. In *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE.

Pradeep Kumar Roy. 2023. A deep ensemble network for sentiment analysis in bi-lingual low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

PC Shilpa, Rissa Shereen, Susmi Jacob, and P Vinod. 2021. Sentiment analysis using deep learning. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 930–937. IEEE.

Thoudam Doren Singh, Telem Joyson Singh, Mirinso Shadang, and Surmila Thokchom. 2021. Review comments of manipuri online video: Good, bad or ugly. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, pages 15–23. Springer.

Srikanth Tammina. 2020. A hybrid learning approach for sentiment classification in telugu language. In *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–6. IEEE.

Farhat Ullah, Xin Chen, Syed Bilal Hussain Shah, Saoucene Mahfoudh, Muhammad Abul Hassan, and Nagham Saeed. 2022. A novel approach for emotion detection and sentiment analysis for low resource urdu language based on cnn-lstm. *Electronics*, 11(24):4096.

V Umarani, A Julian, and J Deepa. 2021. Sentiment analysis using various machine learning and deep learning techniques. *Journal of the Nigerian Society of Physical Sciences*, pages 385–394.

Dang Van Thin, Hung Quoc Ngo, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models. *Journal of Information and Telecommunication*, 7(2):121–143.

Deheng Wang, Aimin Yang, Yongmei Zhou, Fenfang Xie, Zhouhao Ouyang, and Sancheng Peng. 2022. Distillation language adversarial network for cross-lingual sentiment analysis. In *2022 International Conference on Asian Language Processing (IALP)*, pages 45–50. IEEE.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023a. Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *arXiv preprint arXiv:2305.00090*.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023b. Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *arXiv preprint arXiv:2305.00090*.

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. 2022. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. *arXiv preprint arXiv:2211.06607*.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60:617–663.