ISSN 2007-9737

Sentiment Analysis with Khasi Low-Resource Language through Generation of Sentiment Words Using Machine Learning

Banteilang Mukhim*, Arnab Kumar Maji, Sufal Das

Department of Information Technology, North-Eastern Hill University, Shillong, INDIA,
mukhimb@gmail.com, arnab.maji@gmail.com, sufal.das@gmail.com

Abstract. Sentiment Analysis is a Natural Language Processing (NLP) technique to find out the opinion and classify the opinion expressed in a text data with polarity (e.g., positive, negative and neutral). Khasi NLP is just starting to take shape, and ways back as compared to some Indian languages. Sentiment analysis with low resource language is a challenging task as the input data has limited annotated data. The proposed method suggests employing machine translation for the Khasi-English language pair to extract emotion-carrying words from Khasi text using an English emotion word dictionary. Despite the lack of specific sentiment analysis resources for Khasi, this approach enables the identification of sentiment-bearing phrases. After generation of Khasi sentiment words, a transformer-based model is considered for sentiment analysis as a validation tool.

Keywords. Sentiment analysis, Khasi sentiment words, emotion mining, Khasi Language, Khasi sentiment classification.

1 Introduction

Sentiment analysis is the technique of analyzing the input text data into positive, negative, or neutral sentiments. It is the contextual opinion mining of words/sentence that reveals the user's sentiment. The application of sentiment analysis is to evaluate user views to assist in the growth of businesses. It focuses not only on sentiments (positive, negative, and neutral) but also on emotions (happy, sad, angry, etc.) [4]. It employs a variety of NLP tasks, including rule-based analysis, automatic annotation, and hybrid methods. Sentiment words are instrumental in identifying sentiment orientations and intensity [12]. There are two types of individual words: base type words are used to look at regular sentiment and comparative words are used to identify semantically more complex sentiment[11].

These opinions present a way for sentiment analysis, but the complexity to categorize text in their opinions or sentiment is different for different languages [2]. This task mainly relies on the availability of resources for that particular language. For example, the abundance of digital dictionaries, corpora, and labeled datasets advanced the science of sentiment analysis for English quite well, but not as well or as quickly for lower-resource languages such as the Khasi language [13].

Sentiment analysis with low resource language refers to the method to classify the sentiment for the given input text data which is less annotated data as training data [15]. This is a challenging issue in many real-world applications where collecting and annotating large amounts of textual data can be very complex and not efficient [10, 6].

To solve this issue, several techniques have been proposed for sentiment analysis with low resource text data with natural language processing techniques. Several learning methods like transfer learning, multi-task learning, unsupervised learning, active learning etc. are applied [5].

2 Related Works

Substantial research on semantic analysis with low-resource language has been done, for example in Urdu and Bengali. There is very few research works have been performed on

doi: 10.13053/CyS-29-3-5915

any text-processing task with Khasi low-resource language [7].

SentNoB provided a sentiment analysis with noisy Bangala text. In this method, three sentiment labels are categorized from political to agricultural data [9]. BEmoC method was proposed with Bengali Emotion Corpus for emotion detection from Bengali text [8]. Here, six classes of sentiment were considered. These text sentiment classification methods have been served as inspiration for sentiment analysis with low-resource language. Table 1 shows a comparison between a few sentiment analysis research and study for low resource language.

A comparison table of the methods is given in the table 1 for the approaches taken by various research papers for tackling sentiment analysis in low resource languages, these may provide insight on the process to be followed so the task of sentiment analysis can be research and perform for other low resource languages like Khasi.

Several research issues in sentiment analysis with Khasi language are still untouched.

Insufficient Collection of Annotated Khasi Data: A large and high-quality sentiment annotated corpus in Khasi text data is to be created so that this can be used as training data. This will help to enrich sentiment analysis models in this domain.

Lack of Lexical Resources: There is a lack of Khasi language dictionaries, Parts-of-Speech (POS) tagging and other important lexical resources. These resources should be available specifically for the Khasi language to increase the precision of sentiment analysis models.

3 Proposed Methodology

When it comes to digital representation, Khasi, like many other indigenous languages, suffers severe difficulties. Even though Khasi has a rich written tradition, there are not enough digital resources or data to support computational analysis of the language. As a result, sentiment analysis is significantly hampered because machine learning models need a lot of digital text to be trained.

To perform sentiment analysis on Khasi text, the development of a machine translator is proposed for the Khasi-English language pair. The translator

Table 1. Comparison with existing methods

Method	**Dynamic Data**	**Grammar Error Check**	**Low Resource**	**Emoji and Symbols**	**Pre-processing**	**Polar Lexicons**	**Translated Data**	**Labeled Data**
Assamese POS (Part-of-Speech) LEX (Lexical) Analysis [3]	No	✓	✓	×	✓	×	×	✓
Translated Valence Aware Dictionary for sEntiment Reasoner (VADER) [1]	No	✓	✓	✓	✓	✓	×	×
Bengali Cross-Lingual [16]	Yes	×	✓	×	✓	×	✓	✓
Machine Learning Classifier [14]	No	✓	✓	×	✓	×	×	✓
Deep Learning Classifier [14]	No	✓	✓	×	✓	×	×	✓
Verb Based Manipuri Text [15]	No	×	✓	×	✓	×	×	✓

would be developed using whichever method is selected, such as a transformer-based model. Once trained, the translator can be used to extract emotion-carrying words or lexicons from Khasi text using an English emotion word dictionary, such as Valence Aware Dictionary for sEntiment Reasoner (VADER).

The emotion-conveying words obtained from the Khasi text can be translated from English using the translator, and their corresponding emotion weights can be determined using the English dictionary.

The stages followed in this method can be broadly categorized as below

— **Data Collection:** Three types of data are required to be obtained, first - a list of emotional words in the English language, second - a parallel dataset of Khasi and English, to be used for generating a rule, and a translator, and third - a sentiment dataset which in this case will be review text and comments.

— **Association Rule Mining:** Using the rule mining approach generate rules to obtain emotion word pairs and assign emotion valence to the obtained words.

— **Transformer Translator:** Create a sequence-to-sequence transformer that is able to translate Khasi to English, this will be used as a validator.

— **Apply Modified VADER:** Applying the VADER algorithm that is modified for Khasi language to obtain sentiment labels on the Khasi review data.

— **Validation:** Translating the review data to English and applying English VADER to the translated text the sentiment label in English is obtained and these labels are validated that they are within the same polarity, and if so, the predicted sentiment label is correct.

Figure 1 shows a flow chart of the proposed method. The method was split into small tasks, described below with examples is the task performed to realize the proposed method.

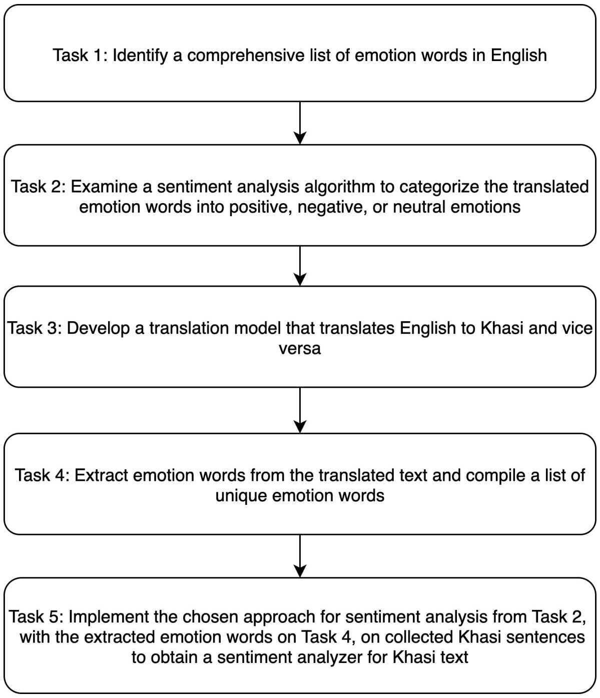
Task 1: Identify a comprehensive list of emotion words in English
Task 2: Examine a sentiment analysis algorithm to categorize the translated emotion words into positive, negative, or neutral emotions
Task 3: Develop a translation model that translates English to Khasi and vice versa
Task 4: Extract emotion words from the translated text and compile a list of unique emotion words
Task 5: Implement the chosen approach for sentiment analysis from Task 2, with the extracted emotion words on Task 4, on collected Khasi sentences to obtain a sentiment analyzer for Khasi text

Fig. 1. Block Diagram of Proposed Methodology

3.1 Collection of Khasi Text Data

Through various online resources, including sentiment analysis dictionaries like VADER, it is possible to obtain an extensive list of emotional words used in the English language. An emotion score is assigned to each English word and phrase in the VADER lexicon, indicating whether it is positive or negative.

Then using a machine translator to translate the Lexicons, terms may be found in the Khasi text that convey emotions before applying the VADER algorithm. The lexicons can then be given sentiment ratings using the VADER dictionary, and those values are utilized to create emotion-weighted lexicons for Khasi sentiment analysis. This method enables performing sentiment analysis on Khasi text even in the absence of complete resources for the Khasi language by utilizing the resources available for English sentiment analysis.

Parallel English-Khasi Data.
These data are collected from the readily available

English-Khasi Bible Translation, along with a similar amount of text data collected from the website http://tatoeba.org/en which is a large database of sentences and translations. These data are used for two cases, one of which will help in generating the rules using rule mining to obtain a word pair of English and Khasi which will then be used to create sentiment lexicons for the Khasi VADER, while the other is to be used for training a transformer translator that will become the English-Khasi translator.

Example:

Table 2. Parallel dataset English-Khasi

Eng	And to rule over the day and over the night, and to divide the light from the darkness: and God saw that it was good.
Kha	ban synshar halor ka sngi bad ka miet bad ban pynkhlad ia ka jingshai na ka jingdum. Te u Blei U la iohi ba ka la long ka ba bha.

3.2 Examine to Categorize the Translated Emotion Words

In the experiment, sentiment analysis of Khasi text was carried out using an adaptation of the VADER algorithm. Based on a vocabulary of terms that have already been rated in English, the computer assigns sentiment values to words in the Khasi text. A list of words in the VADER lexicon is given a sentiment score, ranging from -1 (strongly negative sentiment) to 1 (strongly positive sentiment).

The VADER dictionary consists of the following,

1. Emotion Word/Lexicon: This column contains the list of words that are stored in the lexicon.
2. Valence: This column contains the sentiment score for each word, ranging from -4 to 4. The valence score indicates the degree of how positive or negative associated with the word. A score of 0 indicates neutrality.
3. Probability Score: This column contains a score between -1 and 1, which indicates the probability of the word being used in a positive, negative, or neutral context.
4. List of Intensifiers: Intensifiers are adverbs or adjectives that modify the intensity of the sentiment conveyed by a word. In the context of the VADER dictionary, intensifiers refer to the strength or degree of sentiment conveyed by a particular word. The numbers in the list represent the intensity scores assigned to the word by different raters.

Table 3. Example words from the VADER dictionary

Word	Valence	Positive probability score	Negative probability score
Amazingly	1.9	0.87632	0.12368
Fantastic	2.3	0.93218	0.06782
Terrible	-2.5	0.07023	0.92977
Horrific	-3.8	0.03212	0.96788
Ordinary	0.0	0.5	0.5
Neutral	0.0	0.5	0.5

A modified negation dictionary and booster dictionary have been utilized to modify the method for Khasi. The negation dictionary provides a list of words that negate the sentiment of a previous word, whereas the booster dictionary contains a list of words that amplify or de-intensify sentiment. The addition of Khasi words to these dictionaries made it possible to analyze sentiment in Khasi text more accurately.

3.3 Extracting Khasi Emotion Words Using Rule Mining

This method evaluates the reliability and frequency of translations between source and target words using confidence, support, and count metrics. By splitting the dataset's sentences into individual words, corresponding word pairs are created, as each word in the English half is paired with all khasi words. The counts of these word pairs are then computed to determine their frequency of occurrence. The proportion of sentences containing a certain word pair is calculated to determine its support. And confidence estimates the likelihood of a target word given a source word. High-confidence pairs are selected after sorting the word pairs based on confidence and applying a threshold. To eliminate pairings with insufficient occurrence frequency, a minimum support threshold is selected and those pairs below this threshold are removed. The remaining word pairings constitute the basis for translation rules, with each source word mapping to its associated destination term.

3.4 Develop a Translation Model (Khasi to English).

Initially, a sequence-to-sequence Transformer model was trained using the English and Khasi versions of the accessible Bible dataset, drawing inspiration from other studies on language pairs like English-Spanish and English-French. The model is retrained on the English-Khasi word pair downloaded from http://tatoeba.org/en.

Given the limited dataset available for the English-Khasi language pair, the Transformer model shows satisfactory results and can be used as a translator in most cases. This translator can also translate Khasi to English and apply English VADER on the English text translated from Khasi to compare the result when Modified Khasi VADER is applied on Khasi text and when the sentiment analysis is performed on translated English text.

3.5 Design Sentiment Analysis With the Extracted Emotion Words

To perform sentiment analysis on Khasi text using the VADER algorithm and the extracted Khasi emotion lexicons, the following steps can be followed:

1. Preprocessing: This entails removing unnecessary elements such as punctuation marks, special characters, and numerical digits. Following that, the text can be tokenized into unique words or tokens.
2. Loading Khasi Emotion Lexicons: An important step in this process of sentiment analysis is the successful assimilation of the extracted Khasi emotion lexicons, which encompass those words found within the Khasi text that shows semantic similarity with their English counterparts.
3. Sentiment Intensity Calculation: To analyze the sentiment intensity of individual words in the Khasi text, it becomes necessary to employ the VADER algorithm. This entails assessing the sentiment strength of each word in the context of the Khasi language, guided by the corresponding valence scores obtained from the Khasi emotion lexicons.
4. Polarity Calculation: A positive sum signifies the presence of positive sentiment, whereas a negative sum signifies negative sentiment. Conversely, a sum that approximates to zero signifies a state of neutral sentiment pervading the text.
5. Normalization: To obtain the polarity score for analysis and interpretation, the cumulative sum of the polarity is to be subjected to the process of normalization. Utilizing a normalization equation the score is then normalized in the range of -1 to 1, the polarity score is appropriately scaled, ensuring a standardized evaluation of sentiment intensity.

4 Experiments

4.1 Rule-Based Word Pair Generation

For the English half, as the llist of sentiment words is already available, the sentiment-contributing words are filtered, while for Khasi, POS is used to clean nouns and pronouns as they do not contribute to sentiment. Table 4 shows sample data before and after preprocessing.

After cleaning and tokenizing the sentences, word pairs are generated consisting of each English word matched with all khasi words. This is obtained by looping through the sentences and then for each tokenized word in a sentence, combine it with all possible Khasi words. The result

Table 4. Sample data before and after preprocessing

Before preprocessing

En	And the earth was without form and void, and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters.
Kha	Ka khyndew ka la long kaba khlem dur bad kaba suda, bad ka jingdum ka la tap khlup ïa ka Duriaw bah. U Mynsiem U Blei U la khih halor ki um.

After preprocessing

En	[darkness, spirit, god]
Kha	[long, khlem, dur, suda, jingdum, tap, duriaw, mynsiem, blei, khih]

Table 5. English-Khasi pairs

English Word	**English-Khasi Pairs**
darkness	(darkness, long), (darkness, khlem), (darkness, dur), (darkness, suda), (darkness, jingdum), (darkness, tap), (darkness, duriaw), (darkness, mynsiem), (darkness, blei) (darkness, khih)
spirit	(spirit, long), (spirit, khlem), (spirit, dur), (spirit, suda), (spirit, jingdum), (spirit, tap), (spirit, duriaw) (spirit, mynsiem), (spirit, blei) (spirit, khih)
god	(god, long), (god, khlem), (god, dur), (god, suda), (god, jingdum), (god, tap), (god, duriaw) (god, mynsiem), (god, blei), (god, khih)

of this process is depicted in the Table 5. In other words, for one sentence pair multiple word pairs are obtained and further along they are filtered using rule mining metrics to obtained the pair that is most likely.

Next, to obtain the rules, a rule mining method following apriori rule mining is used to calculate three metrics for each of the word pair - count, support, and confidence. Taking the word pair dataset, the frequency occurrence of each word pair is calculated, which is done by simply counting the number of the word pairs. An example is shown in Table 6. In the example, it can be seen that for each English word with its word pair, there are some counts which are higher. Moreover, between different English words, it is seen that there is a disparity between the count range. This is due to the different number of word pair occurrences in the sentences. Word pairs that are below a predefined minimum threshold are removed.This reduces the number of word pairs and eliminates any word pairs that do not contribute to the rule. A different threshold is set for each source word which is calculated by finding the average count of the word pairs for a particular source word and adding an offset based on the requirement. Next, the support of each word pair is calculated which is obtained by dividing the count of the word pair by the number of sentences in the dataset. In this work, the count is divided by the total number of sentence pair in the dataset. Table 7 shows a few examples of the word count with its support.

Similarly, a threshold for the support is also set, the word pairs below this threshold are removed. This threshold is calculated by getting the average and adding an offset based on the requirement for word pairs of each source word. Next, confidence is calculated, which, as the name suggests, indicates how likely it is that the word pair rule is correct. The higher the value, the greater the confidence. To calculate confidence, divide the word pair count by the count of the source word (which in this case is the English word). Then confidence in English to Khasi is obtained. For Khasi to English, divide it by the count of the Khasi word, as Khasi becomes the source word. An example is shown of a few word pairs and the confidence in Table 8.

The last step is to create the rules where the rule is selected from the final word pair. If for one source word, there are multiple word pair still,

Table 6. Example: Word pairs with their count

Word pair	Count
(darkness, suda)	1
(darkness, jingdum)	10
(darkness, duriaw)	2
(spirit, mynsiem)	57
(spirit, blei)	18
(spirit, khih)	1
(god, mynsiem)	33
(god, blei)	1089
(god, khih)	4

Table 7. Example: Word pairs with their count and support

Word pair	Count	Support
(darkness, jingdum)	10	0.00213
(spirit, mynsiem)	57	0.01217
(spirit, blei)	18	0.00384
(god, mynsiem)	33	0.00704
(god, blei)	1089	0.23264

select the word pair with the highest confidence and support. The resulting word pair will then be the rule for that source word.

From the above example, the rules are: (darkness → jingdum), (spirit → mynsiem), (god → blei).

4.2 Modified VADER for Khasi Language

VADER is a text sentiment analysis model leveraging a lexical approach, that is sensitive to both emotion polarity (positive/negative) and intensity. VADER's aim is to map words to emotions by creating a lexicon or 'dictionary of sentiment'. This dictionary can be used to assess the sentiment of phrases and sentences without having to look at anything else. VADER sentiment analysis is based on a dictionary that maps lexical features into emotion intensities known as sentiment scores. A text's sentiment score can be calculated by adding the intensity of each word in the text. The strength of emotions, or sentiment score, is evaluated on a scale of -4 to +4, with -4 being the most negative and +4 being the most positive. The midpoint 0 denotes neutral sentiment.

Table 8. Example: Word pairs with their count, support and confidence

Word pair	Count	Support	Confidence
(darkness, jingdum)	10	0.00213	0.901
(spirit, mynsiem)	57	0.01217	1.000
(god, blei)	1089	0.23264	1.000

The sentiment score of a sentence is calculated by summing up the sentiment scores of each VADER-dictionary-listed word in the sentence and then normalization is applied to the total to map it to a value between -1 to 1. Normalization is done by the formula:

$$NormalizedScore = \frac{x}{\sqrt{x^2 + \alpha}}. \quad (1)$$

Here x is the sum of all the valence scores of the emotion words in the sentences and α is set to 15 and is the assumed maximum for an English valence score. Moreover, the sentiment of the sentence is also dependent upon punctuation, capitalization, modifiers, and negations.
A dictionary is created for a few selected Khasi text of known polarity, and the result observed is shown in Table 9.

Table 9. Few Sample Khasi VADER LEX Polarity Analysis

Text	Translated Khasi VADER	Known Polarity
'ba bha bad ba sniew' (its good and bad)	-0.1531	0
'bym bha bad bym sniew' (not good or bad)	0.1139	0
'ka sorkar jong ngi ka iarap' (our government helps)	0.4019	1
'ka sorkar jong ngi bym iarap' (our government dosent help)	-0.3089	-1

VADER analysis also takes this into account during sentiment analysis. In this work, the English

VADER is modified to adapt to the Khasi language by modifying the sentiment dictionary, the booster dictionary, and the negation dictionary, while the core functionality of the algorithm remains the same. The sentiment words for the dictionary are available from the rule generated for translation, and the sentiment value for the Khasi sentiment word is copied over from the corresponding English word. A sample of the sentiment dictionary for the modified Khasi VADER is shown in Table 10.

Table 10. Khasi Sentiment Dictionary for Modified VADER

Khasi Sentiment Lexicon	English Equivalent	Valence
jingdum	darkness	-1.0
mynsiem	spirit	0.7
blei	god	1.1
mynjur	agree	1.5
lyngngoh	shock	-1.6
mraw	slavery	-3.8
thaba	bright	1.9

To obtain the negation words, adverbs that occur with the emotion words are filtered, and cross-checked manually with the help of the rule generated as explained in section 4.1. These negation words will help identify sentences that contain words that flip the polarity of the sentence, and identify negative sentences.

Using the modified VADER on Khasi sentences the sentiment score is obtained by summing the valence of the sentiment value of the sentiment word present in the sentence and then normalize it in the range of -1 to 1. Shown in Table 11 is an example of the result produced by the modified VADER, adapted for the Khasi language.

The analysis performed is based only on the sentiment lexicons and the negation dictionary. Although the result is good, to further increase the accuracy and to compensate for subjectivity, the calculated score can also be subjected to a booster dictionary, also adding an exhaustive list of Khasi idioms to find commonly used phrases and slangs that may enhance or decrease the overall sentiment score of the sentence.

Table 11. Example of VADER Output

English: And the earth was without form and void, and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. **Res :** {'neg': 0.063, 'neu': 0.818, 'pos': 0.119, 'compound': 0.2023}
Khasi: Ka khyndew ka la long kaba khlem dur bad kaba suda, bad ka jingdum ka la tap khlup ia ka Duriaw bah. U Mynsiem U Blei U la khih halor ki um. **Res :** {'neg': 0.057, 'neu': 0.833, 'pos': 0.109, 'compound': 0.2023}

In the example Table 11 the English version of the sentence is analysed using nltk VADER algorithm, and for the Khasi version a modified version of the same algorithm excluding the idiom checking and adverb booster is applied, providing a result that is near the same sentiment value for both version of the same sentence.

4.3 Translator for Khasi Language Using Transformer

A transformer model for translation of the English to Khasi text and vice versa, is developed for the purpose of validating the result of the sentiment score output by the modified Khasi VADER. The transformer translator translates the sentiment dataset from Khasi language to English. The English VADER algorithm is then run on the translated text, and if the output of this is comparable with the output of the Khasi VADER, it indicates the confidence of the algorithm. Creation of the Transformer follows the guided transformer for other language pairs with some modifications to adapt to this specific work. The steps are detailed as follows:

Preparation and Data Loading:

Starts by utilizing the necessary libraries for numerical operations and the required modules from TensorFlow and Keras for building and training the Transformer model. Additionally, modules for text preprocessing are imported. Then the parallel dataset of English-Khasi sentence pairs is loaded.

Data Preparation and Data Splitting:

A function to create text pair that takes an English sentence and a Khasi sentence and formats them into a text pair of tuples for better processing is defined. The function adds special <start> and <end> tokens to the Khasi sentence to mark the beginning and end. The function is applied to each row of the English-Khasi data, resulting in a list of text pairs.

Using the *train_test_split* function from *sklearn.model_selection* to split the text-pairs into training and validation sets.

Text Pre-processing:

The next step is to preprocess and clean the data from special characters and punctuations by removing them and setting the sentence to lowercase. The characters that need to be removed from the text during preprocessing, include punctuation marks and special characters.

Text Vectorization:

Two TextVectorization objects are created: English Vectorization and Khasi Vectorization. These objects are configured with the specified vocabulary size, output mode, and sequence length. A custom standardization function is used for text standardization, which converts text to lowercase and removes specified characters. Then on the vectorization objects learn the vocabulary from the training data into the vectorization objects.

Training Data Preparation:

Before training the data is prepared in a format acceptable as input to the transformer. This involves tokenizing the English and Khasi sentences, truncating or padding them to the specified sequence length. A TensorFlow Dataset is then created from the pairs of English and Khasi sentences. The data is batched, mapped with the format data function, shuffled, and cached for better performance.

Model Building:

The transformer architecture is used for the model. The architecture includes the transformer encoder which performs self-attention and feed-forward operations on the input sequence. The Positional Embedding layer is used to incorporate positional information into the input embedding. The Transformer Decoder layer performs self-attention and attention over the encoder outputs. The decoder side of the transformer decodes the encoder outputs which, in effect, translates Khasi to English words.

Model Training:

The transformer model is trained using the rmsprop (Root Mean Squared Propagation) optimizer and sparse categorical cross-entropy loss. The fit function is called to train the model on the training dataset, with a specified number of epochs. The validation dataset is used for validation during training.

Model Evaluation:

After training, the vocabulary and index lookup for the Khasi language is prepared. A decode sequence function is defined to generate translations for English sentences. A random text pair from the test set is selected, and the decode sequence function is called to translate the English sentence to Khasi.

The model is trained and tested for results in a Python notebook, the Table 12 shows an example of the translation performed.

5 Results Analysis

5.1 Comparison of Valence Score

A few random data from a list of sentences are selected and Khasi VADER analysis is performed, the sentences are then translated to English and then the English version of VADER is applied, the sentiment score value is then compared in a chart shown in Figure 2.

It can be seen from the chart that all the bars for English and Khasi score lies on the same side, this infers that the modified VADER algorithm is predicting the sentiment score correctly. The length of the graph shows the difference in the score. In a few cases the sentiment score intensity is not predicted correctly.

Table 12. Example of Transformer Translation

English	And they said, We will call the damsel,and inquire at her mouth.
Khasi	Ki la ong, “To ngin khot ïa ka samla bad ngin kylli ïa ka.”
Translated	ki la ong to ngin khot ïa ka samla bad ngin kylli ïa ka

English	So he went with them. And when they came to Jordan, they cut down wood.
Khasi	Kumta u la leit lang bad ki. Haba ki la poi ha Jordan, ki la iapom dieng.
Translated	kumta la leit lang bad ki haba la poi ha jordan ki la iapom dieng.

5.2 Performance of Proposed Method

Utilizing a labeled dataset of sentiment Khasi text data, a comparison of performance between the modified VADER algorithm and its English counterpart, on the translated text, is tabulated as shown in Tables 13 and 14.

Table 13. Classification Report for khasi text modified khasi VADER

Khasi	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	0.79	0.60	0.68	177
Neutral	0.62	0.66	0.64	139
Positive	0.64	0.70	0.70	181
Accuracy			0.67	497
Macro Avg	0.69	0.68	0.67	497
Weighted Avg	0.73	0.72	0.72	497

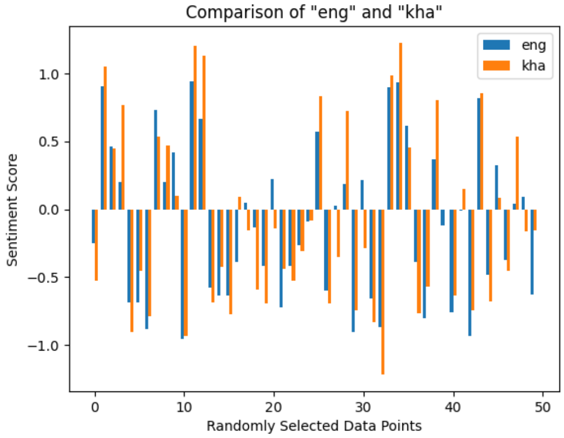

Fig. 2. the Bar chart of Comparison of Sentiment Scores in Khasi and English Vader

Table 14. Classification Report for English version of same text and analysis using English VADER

English	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	0.84	0.64	0.72	177
Neutral	0.66	0.70	0.68	139
Positive	0.68	0.81	0.74	181
Accuracy			0.72	497
Macro Avg	0.73	0.71	0.71	497
Weighted Avg	0.73	0.72	0.72	497

5.3 Comparison with Existing Methods

The method of obtaining a lexicon using a modified VADER for Khasi sentiment analysis has been found to be better as it produces better accuracy compared to other methods of sentiment classification. The lack of preprocessing tools with other methods, lexicon-based analysis such as VADER provides better classification, as depicted by the results in Table 15. As in the table, the proposed VADER method shows an accuracy that is equivalent to that obtained by SVM (support vector machine). This comparison was done on a common English dataset. A visual comparison is shown in figure 3.

6 Conclusion

In general, it may be claimed that the sentiment analysis for Khasi text has been performed to an

Table 15. Performance Comparison of Classifiers

Classifier	ACC	Precision	Recall	F1-Score
Proposed Method	**0.67**	**0.716**	**0.716**	**0.719**
Naive-Bayes	0.65	0.683	0.683	0.683
TF-IDF	0.60	0.650	0.650	0.650
SVM	0.66	0.727	0.710	0.718

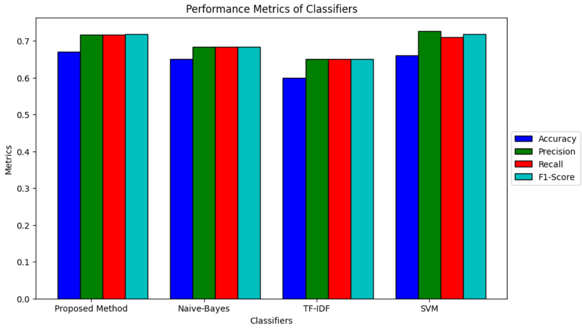

Fig. 3. Performorrmance Comparison

acceptable degree of performance. The use of a standard, tried-and-true method for accomplishing the task has proven successful, and with the help of a relatively modern tool like a Transformer, the method has been supported as a validation tool.

In the method developed for completing the sentiment analysis task, the use of the VADER algorithm has not only provided an overall sentiment classification into one label but also offered insight into the probability of the sentiment belonging to other labels. In this case, there are three labels: negative, neutral, and positive. This probability reflects the subjectivity of the sentiment and indicates how subjective the sentiment is in relation to the assigned polarity score. As for using apriori rule-based mining, which may be considered a relatively less sophisticated and older data mining concept compared to current technologies, it remains effective given the limited resources available for NLP in the Khasi language. This method, which requires little to no tools for finding data patterns, yields accurate results. Although there is room for improvement, what has been achieved is satisfactory as proof of concept, demonstrating that the method works and provides good results.Utilizing the VADER algorithm allows for sentiment analysis in Khasi text using a well-established and widely used sentiment analysis approach. The extracted Khasi emotion lexicons provide a foundation for sentiment analysis, enabling sentiment scoring based on the lexicons' valence scores.

Collecting and expanding the dataset of Khasi emotion lexicons can help improve the accuracy and coverage of sentiment analysis in Khasi. Refining the translation model used to extract Khasi emotion lexicons can enhance the accuracy of the sentiment analysis process. Considering domain-specific sentiment lexicons for Khasi text can help improve the sentiment analysis results for specific domains or topics.

References

1. **Amin, A., Hossain, I., Akther, A., Alam, K. M. (2019).** Bengali vader: A sentiment analysis approach using modified vader. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp. 1–6.

2. **Bai S, P., Kumar G K, R. (2016).** Efficient incremental itemset tree for approximate frequent itemset mining on data stream. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 239–242. DOI: `10.1109/ICATCCT.2016.7912000`.

3. **Das, R., Singh, T. D. (2021).** A step towards sentiment analysis of assamese news articles using lexical features. Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India, Springer, pp. 15–23.

4. **Das, S., Kalita, H. K. (2018).** S3call: A semi-supervised sentiment classification method for large web-based data. Journal of Computational and Theoretical Nanoscience, Vol. 15, No. 6-7, pp. 2264–2268.

5. **EL-Haj, M., Kruschwitz, U., Fox, C. (2014).** Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. Language Resources and Evaluation.

6. **Gangula, R. R., Mamidi, R. (2018).** Resource creation towards automated sentiment analysis in Telugu (a low resource language) and integrating multiple domain sources to enhance sentiment prediction.

7. **Haddi, E., Liu, X., Shi, Y. (2013).** The role of text pre-processing in sentiment analysis. Procedia Computer Science, Vol. 17, pp. 26–32. DOI: 10.1016/j.procs.2013.05.005.

8. **Iqbal, M. A., Das, A., Sharif, O., Hoque, M. M., Sarker, I. H. (2022).** Bemoc: A corpus for identifying emotion in bengali texts. SN Computer Science, Vol. 3, No. 2, pp. 135.

9. **Islam, K. I., Kar, S., Islam, M. S., Amin, M. R. (2021).** Sentnob: A dataset for analysing sentiment on noisy bangla texts. Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3265–3271.

10. **Kaur, G., Kaur, K. (2015).** Sentiment analysis on Punjabi news articles using SVM. Int J Sci Res, Vol. 6, No. 8, pp. 414–421.

11. **Liu, B. (2015).** Sentiment analysis: Mining opinions, sentiments, and emotions. DOI: 10.1017/CBO9781139084789.

12. **Liu, B. (2022).** Sentiment analysis and opinion mining. Springer Nature.

13. **Mahmoud, H. A. H., Mengash, H. A. (2021).** Machine translation utilizing the frequent-item set concept. Sensors, Vol. 21, No. 4. DOI: 10.3390/s21041493.

14. **Meetei, L., Singh, T. D., Borgohain, S., Bandyopadhyay, S. (2021).** Low resource language specific pre-processing and features for sentiment analysis task. Language Resources and Evaluation, Vol. 55. DOI: 10.1007/s10579-021-09541-9.

15. **Nongmeikapam, K., Khangembam, D., Hemkumar, W., Khuraijam, S., Bandyopadhyay, S. (2014).** Verb based manipuri sentiment analysis. Special Issue on NLPACC, International Journal on Natural Language Computing (IJNLC), Vol. 3. DOI: 10.5121/ijnlc.2014.3311.

16. **Sazzed, S. (2020).** Cross-lingual sentiment analysis in bengali utilizing a new benchmark corpus. Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generate, pp. 50–60.

Article received on 16/04/2025; accepted on 22/07/2025.
**Corresponding author is Banteilang Mukhim.*